Linear Algebra and Multivariable Calculus

Notes from 18.02 Fall 2024

Evan Chen

16 October 2024

Table of contents

§1 Preface

At MIT, the course 18.02 (multivariable calculus) is a general institute requirement (GIR); every student must pass this class in order to graduate. These are lecture notes based upon the fall 2024 instance of the course, taught by Davesh Maulik.

§1.1 [TEXT] Goals of this book

These notes have the following lofty goal:

> **Goal**
>
> In theory, an incoming MIT student with a single-variable calculus background should be able to pass the 18.02 final exam by **only** reading these notes and problems, working through several practice final exams, and going to a weekly office-hours[1] to ask questions to a real human.

This is ambitious, and your mileage my vary. Just to be clear, this text is unofficial material and there is no warranty or promise that this goal will be fulfilled for you. (Also, if you are actually an MIT student, bear in mind the content of the course will vary by instructor.) But with this goal in mind, here are some parts of the design philosophy of this book.

- **It's practical**. It sticks to the basics and emphasizes giving straight cookbook-like answers to common exam questions.
  - I better say something about memorizing recipes. In principle, if you have perfect memory, you could potentially get a passing score (but not a perfect score) on the final exam by *only* memorizing the recipes.

    I don't recommend this approach; even a vague conceptual understanding of where a recipe is at minimum quite helpful for remembering said recipe. But it may be useful to know in principle that the recipe is all you need, and conversely, that you should have the recipes down by heart.
- **It's concrete**. We only work in $\mathbb{R}^n$, and not a generic vector space. We don't use anywhere near the level of abstraction as, say, the Napkin[2]. We don't assume proof experience.
- **It writes things out**. A lot of lecture notes were meant to accompany a in-person lecture rather than replace it. These notes are different. They are meant to stand alone, and anything that would normally be said out loud is instead written out in text.
- **It has full solutions to its exercises**. I really believe in writing things out. I'd rather have a small number of exercises with properly documented solutions than an enormous pile of mass-produced questions with no corresponding solutions.

  > **TODO**: Okay this is not true yet lol I'm working on it. There will be solutions one day. Especially since ChatGPT can do all the exercises anyway kappa.

- **It tries to explain where formulas come from**. For example, these notes spell out how matrix multiplication corresponds to function composition (in Section 7.3), something that isn't clearly stated in many places. I believe that seeing this context makes it easier to internalize the material.

[1] You can substitute the office hours for a knowledgeable friend, or similar. The point is that you should have at least some access to live Q/A.

- **It marks more complicated explanations as "not for exam"**. I hope the digressions are interesting to you (or I wouldn't have written them). But I want to draw a clear boundary between "this explanation is meant for your curiosity or to show where this formula comes from" versus "this is something you should know by heart to answer exam questions".

  There are two kinds of ways we mark things as not for exam:

  - Anything in a gray digression box is not for exam.

    > **Digression**
    >
    > Here's an example of a digression box.

  - Anything in an entire section marked **[SIDENOTE]** is not for exam.

- **It's written by Evan Chen**. That's either really good or really bad, depending on your tastes. If you've ever seen me teach a class in person, you know what I mean.

§1.2 [TEXT] Prerequisites

As far as prerequisites go, a working knowledge of pre-calculus and calculus as taught in United States high schools is assumed.

- **Algebra**: You should be able to work with elementary algebra, so that the following statements make sense

  $$x^2 - 7x + 12 = (x-3)(x-4) = 0 \implies x = 3 \quad \text{or } x = 4.$$

  You should also be able to solve two-variable systems of equations, such as

  $$\begin{cases} 5x - 2y = 8 \\ 3x + 10y = 16 \end{cases} \implies (x, y) = (2, 1).$$

- **Trigonometry**: You should be know how sin and cos work, in both degrees and radians. So you should know $\sin(30°) = \frac{1}{2}$, and $\cos\left(\frac{7\pi}{6}\right) = -\frac{\sqrt{3}}{2}$.

- **Precalculus**: You should know some common formulas covered in precalculus for vectors and matrices:

  - You should be able to add and scale vectors, like

    $$\begin{pmatrix} 1 \\ 7 \end{pmatrix} + 10\begin{pmatrix} 3 \\ 5 \end{pmatrix} = \begin{pmatrix} 1 \\ 7 \end{pmatrix} + \begin{pmatrix} 30 \\ 50 \end{pmatrix} = \begin{pmatrix} 31 \\ 57 \end{pmatrix}.$$

    (It's really as easy as the equation above makes it look: do everything componentwise.)

  - You should know the rule for matrix multiplication, so that for example you could carry out the calculation

    $$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}\begin{pmatrix} 7 \\ 8 \\ 9 \end{pmatrix} = \begin{pmatrix} 1 \cdot 7 + 2 \cdot 8 + 3 \cdot 9 \\ 4 \cdot 7 + 5 \cdot 8 + 6 \cdot 9 \end{pmatrix} = \begin{pmatrix} 50 \\ 122 \end{pmatrix}.$$

    If you haven't seen this before, there are plenty of tutorials online; find any of them. Poonen's notes (mentioned later) do cover this for example; see section 1-2 of `https://math.mit.edu/~poonen/notes02.pdf`.

---

[2]That's the one at `https://web.evanchen.cc/napkin.html`, which *does* assume a proof-based background.

You are *not* expected to have any idea why the heck the rule is defined this way; an explanation for where this rule comes from is in Section 7.3. we will explain what this rule means later. So we'll assume you have memorized this strange rule, but don't know what it means.

- We'll assume you know the formula for the determinant of a 2×2 and 3×3 matrix; that is

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$$

and

$$\det \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{pmatrix} = a_1 \det \begin{pmatrix} b_2 & b_3 \\ c_2 & c_3 \end{pmatrix} - a_2 \det \begin{pmatrix} b_1 & b_3 \\ c_1 & c_3 \end{pmatrix} + a_3 \det \begin{pmatrix} b_1 & b_2 \\ c_1 & c_2 \end{pmatrix}.$$

For example, you should be able to verify the correctness of the following equation:

$$\det \begin{pmatrix} 0 & 1 & 5 \\ 2 & 0 & 13 \\ 1 & 4 & 1 \end{pmatrix} = 51.$$

We won't assume you know where this formula comes from, and in fact we won't be able to explain that within these notes. But if you're curious, you should read Chapter 12 of the Napkin.

- **Calculus**: You should know single variable derivatives and integrals, for example:
  - You should be able to differentiate $x^7 + \sin(x)$ to get $7x^6 + \cos(x)$.
  - You should be able to integrate $\int_0^1 x^2 \, dx$ to get $\frac{1}{3}$.

  This is covered in the course 18.01 at MIT, and also in the AP calculus courses in the United States.

> 🔥 **Tip**
>
> If you're not at MIT, you should replace the words "18.01" and "18.02" with the course names corresponding to "single-variable calculus" and "multi-variable calculus" at your home institution.

No proof-based background is expected.

§1.3 [TEXT] Topics covered

Here is a brief overview of what happens in these parts

Alfa and Bravo This part is dedicated to **linear algebra** (vectors and matrices). This is intentional, because some working knowledge of linear algebra is important. In fact, if I was designing a serious course in multivariable calculus for math majors, it would come after an entire semester of properly-done linear algebra first.

Charlie This short part is review of the **complex numbers** $\mathbb{C}$. I actually don't know why this is part of 18.02, to be honest, but since it happened I included a short section on it.

Delta Covers the calculus of functions $\mathbf{r} : \mathbb{R} \to \mathbb{R}^n$, which is usually thought of as a **parametric** function $\mathbf{r}(t)$ (a time-indexed trajectory through the vector space $\mathbb{R}^n$). This section turns out to be easy because it's pretty much all 18.01 material. This part is therefore also only a few pages long.

Echo and Foxtrot Cover the **differentiation of multivariable functions** $f : \mathbb{R}^n \to \mathbb{R}$, and the optimization such functions. The start of these two parts in the gradient ∇f This is the first serious multivariable calculus usage.

- **TODO**: finish writing this

(The words Alfa, Bravo, Charlie, etc. are from the NATO phonetic alphabet which the author of this book has memorized from overexposure to puzzle hunts.)

§1.4 [TEXT] The structure of this book

You will quickly notice that all the subsections are labeled with different headings. Here's an explanation of what they mean.

TEXT Good old prose. An explanation like you might hear in a lecture.
RECIPE Contains only the final recipe, as you need it on the exam. As I mentioned before, I don't like the idea of just memorizing recipes, but in theory you might still be able to pass the exams by doing only this.
SIDENOTE An optional extended discussion. You can skip these unless you're interested in them.
RECAP A summary of what happened in this chapter.
EXER Problems to work on. Starred exercises are harder than questions that will appear in the actual MIT course.

You'll also see some colored boxes that mark where certain chunks begin and end. These should be self-explanatory.

§1.5 [TEXT] Other references

The best resource I have for 18.02 in text is definitely Bjorn Poonen's fall 2021 notes, available at

`https://math.mit.edu/~poonen/notes02.pdf`.

Poonen is a really great writer of mathematical exposition in general, and I highly recommend these notes as a result. In fact, I will even tell you, for each section, what the corresponding sections of Poonen are if you decide something I write doesn't make sense and you want to reference the corresponding text. (That said, this text is meant to stand alone.)

There are lots of other resources on multivariable calculus out there too. For example, MIT OpenCourseWare has some supplementary notes and problems still used by the math department. And so on.

§1.6 [SIDENOTE] If you're thinking of becoming a math major

If you're thinking of becoming a math major, there's some advice in Section 27.1.

§1.7 [SIDENOTE] Acknowledgments

- Thank you to the staff and other recitation leaders who made this course possible; particularly Davesh Maulik for leading the instance of the course this year full-heartedly and Karol Bacik for making so much happen behind the scenes. Thanks also to Sefanya Hope for coordinating many other logistics, and particularly for helping me book classrooms on short notice on multiple occasions.
- Thank you to all the students in my recitation section (and those officially enrolled in other sections, but who came to my section anyway!) who regularly attended my class every Monday and Wednesday at 9am. That's some real early-morning dedication. There's a saying that the enthusiasm of an instructor can be contagious, but I definitely think the enthusiasm of students can be as well.

  - In particular, I got a lot of words of thanks and encouragements from my students this year, which I am indeed grateful for. I certainly wouldn't have had the motivation to type these notes without these kind words.
- The author thanks Ritwin Narra and Royce Yao for several corrections.
- Finally, the author thanks OpenAI for being gifted a Plus subscription to ChatGPT, which helped a lot with generating sample questions and solutions throughout the document.

TODO: more to come

§2 Type safety

Before we get started with the linear algebra and calculus, I want to talk quickly about *types of objects*. This is an important safeguard for the future in checking your work and auditing your understanding of a topic; a good instructor will point out, in your work, any time you make a type-error.

§2.1 [TEXT] Type errors

In mathematics, statements are usually either true or false. Examples of false statements[3] include

$$\pi = \frac{16}{5} \qquad \text{or} \qquad 2 + 2 = 5.$$

However, it's possible to write statements that are not merely false, but not even "grammatically correct", such as the nonsense equations

$$\pi = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \begin{pmatrix} 0 & 1 & 2 \\ 3 & 4 & 5 \end{pmatrix} = \cos\begin{pmatrix} 6 \\ 7 \end{pmatrix}, \qquad \det\begin{pmatrix} 5 \\ 11 \end{pmatrix} \neq \sqrt{2}.$$

To call these equations false is misleading. If your friend asked you whether $2 + 2 = 5$, you would say "no". But if your friend asked whether π equals the 2×2 identity matrix, the answer is a different kind of "no"; really, it's "your question makes no sense".

These three examples are **type errors**. This term comes from programming: most programming languages have different data types like integer, boolean, string, array, etc., and will usually[4] prevent you from doing anything idiotic like adding a string to an array.

Objects in mathematics work in a really similar way. In the first weeks of 18.02, you will meet real numbers, vectors, and matrices; these are all different types of objects, and certain operations are defined (aka "allowed") or undefined (aka "not allowed") depending on the underlying types. Table Table 1 lists some common examples with vectors you've seen from precalculus.

Operation	**Notation**	**Input 1**	**Input 2**	**Output**
Add/subtract	$a \pm b$	Scalar	Scalar	Scalar
Add/subtract	$\mathbf{v} \pm \mathbf{w}$	Length d vector	Length d vector	Length d vector
Add/subtract	$M \pm N$	$m \times n$ matrix	$m \times n$ matrix	$m \times n$ matrix
Multiply	$c\mathbf{v}$	Scalar	Length d vector	Length d vector
Multiply	ab	Scalar	Scalar	Scalar
Multiply	MN	$m \times n$ matrix	$n \times p$ matrix	$m \times p$ matrix
Dot product	$\mathbf{v} \cdot \mathbf{w}$	Length d vector	Length d vector	Scalar
Cross product	$\mathbf{v} \times \mathbf{w}$	Length 3 vector	Length 3 vector	Length 3 vector
Length/mag.	$\lvert\mathbf{v}\rvert$	Any vector	*n/a*	Scalar
Determinant	$\det A$	Any square matrix	*n/a*	Scalar

Table 1: Common linear algebra operations. For 18.02, "scalar" and "real number" are synonyms.

[3] Indiana Pi bill and 1984, respectively.

[4] JavaScript is a notable exception. In JavaScript, you may know that `[]` and `{}` are an empty array and an empty object, respectively. Then `[]+[]` is the empty string, `[]+{}` is the string `'[object Object]'`, `{}+[]` is `0`, and `{}+{}` is `NaN` (not a number).

> **Digression**
>
> A common question at this point is how you are supposed to figure out whether a certain operation is allowed or not. For example, many students want to try and multiply two vectors together component-wise; why is
>
> $$\begin{pmatrix} 2 \\ 3 \end{pmatrix} \begin{pmatrix} 4 \\ 5 \end{pmatrix} \stackrel{?}{=} \begin{pmatrix} 8 \\ 15 \end{pmatrix}$$
>
> not a legal sentence? It seems like it would make sense.
>
> The answer is that you *don't* have to figure out — you are *told*; Table 1 isn't something that you derive. That is, Table 1 consists of the *definitions* which you have been given.
>
> (Or more sarcastically, it's all just a social construct. Well, it's bit more nuanced than that. Definitions aren't judged by "correctness"; that doesn't make sense; you are allowed to make up whatever definitions you want. Instead, definitions are judged by whether they are *useful*. Which is obviously subjective, but it's less subjective than you might guess.)

§2.2 [TEXT] Why you should care

There are two action items to take away from this section.

§2.2.1 When learning a new object, examine its types first

What this means is that, every time you encounter a new kind of mathematical object or operation (e.g. partial derivative), **the first thing you should do is ask what types are at play**. This helps give you a sanity check on your understanding of the new concept.

We'll use boxes like this throughout the box to do this:

> **Type signature**
>
> This is an example of a type signature box. When we want to make comments about the types of new objects, we'll put them in boxes like this.

§2.2.2 Whenever writing an equation, make sure the types check out

Practically, what's really useful is that if you have a good handle on types, then it **gives you a way to type-check your work**. This is the analog of dimensional analysis from physics, where you know you messed up if some equation has $\mathrm{kg} \cdot \mathrm{meters} \cdot \mathrm{seconds}^{-2}$ on the left but $\mathrm{kg} \cdot \mathrm{meters} \cdot \mathrm{seconds}^{-1}$ on the right.

For example, if you are reading your work and you see something like

$$|\vec{v} \times \vec{p}| = 9\vec{p} \tag{1}$$

then you can immediately tell that there's a mistake, because the two sides are incompatible — the left-hand side is a real number (scalar), but the right-hand side is a vector.

§2.3 [RECAP] Takeaways from type safety

- Throughout this book, every time you meet a new operation, make sure you know what types of objects it takes as input and which it takes as output.

- Whenever you write an equation, make sure it passes a type-check. You can catch a lot of errors like Equation 1 using type safety alone.

§2.4 [EXER] Practice with type safety

TODO: have a list of equations here, and ask to identify the type errors? or similar

Part Alfa: Linear Algebra of Vectors

For comparison, this part corresponds approximately to §1, §2, §3.9 of Poonen's notes.

§3 Review of vectors

§3.1 [TEXT] Notation for scalars, vectors, points

If you haven't seen $\mathbb{R}$ before, let's introduce it now:

Definition

We denote by $\mathbb{R}$ the real numbers, so $3, \sqrt{2}, -\pi$ are elements of $\mathbb{R}$. Sometimes we'll also refer to a real number as a **scalar**.

The symbol "$\in$", if you haven't seen it before, means "is a member of". So $3 \in \mathbb{R}$ is the statement "3 is a real number". Or $x \in \mathbb{R}$ means that x is a real number.

Unfortunately, right off the bat I have to mention that the notation $\mathbb{R}^n$ could mean two things:

Definition

By $\mathbb{R}^n$ we could mean one of two things, depending on context:

- The vectors of length n, e.g. the vector $\binom{\pi}{5}$ is a vector in $\mathbb{R}^2$.
- The points in n-dimensional space, e.g. $\left(\sqrt{2}, 7\right)$ is a point in $\mathbb{R}^2$.

To work around the awkwardness of $\mathbb{R}^n$ meaning two possible things, this book will adopt the following conventions for variable names:

Type signature

- Bold lowercase letters like $\mathbf{u}$ and $\mathbf{v}$ will be used for vectors. When we draw pictures of vectors, we always draw them as *arrows*.
- Capital letters like P and Q are used for points. When we draw pictures of points, we always draw them as *dots*.

We'll also use different notation for actual elements:

Type signature

- A vector will either be written in column format like $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$, or with angle brackets as $\langle 1, 2, 3 \rangle$ if the column format is too tall to fit.
- But a point will always be written with parentheses like $(1, 2, 3)$.

Some vectors in $\mathbb{R}^3$ are special enough to get their own shorthand. (The notation "$:=$" means "is defined as".)

Definition

When working in $\mathbb{R}^2$, we define

$$\mathbf{e}_1 := \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \mathbf{e}_2 := \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

and

$$\mathbf{0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Definition

When working in $\mathbb{R}^3$, we define

$$\mathbf{e}_1 := \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{e}_2 := \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{e}_3 := \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

We also let

$$\mathbf{0} := \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

In other places, you'll sometimes see $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ instead, but this book will always use $\mathbf{e}_i$.

§3.2 [TEXT] Length

Definition

The **length** of a vector is denoted by $|\mathbf{v}|$ and corresponds to the length of the arrow drawn. It is given by the Pythagorean theorem.

- In two dimensions:

$$\mathbf{v} = \begin{pmatrix} x \\ y \end{pmatrix} \Longrightarrow |\mathbf{v}| := \sqrt{x^2 + y^2}.$$

- If three dimensions:

$$\mathbf{v} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \Longrightarrow |\mathbf{v}| := \sqrt{x^2 + y^2 + z^2}.$$

In n dimensions, if $\mathbf{v} = \langle x_1, ..., x_n \rangle$, the length is $|\mathbf{v}| := \sqrt{x_1^2 + ... + x_n^2}$.

Type signature

The length $|\mathbf{v}|$ has type scalar. It is always positive unless $\mathbf{v} = \mathbf{0}$, in which case the length is 0.

§3.3 [TEXT] Directions and unit vectors

Remember that a vector always has

- both a **magnitude**, which is how long the arrow is in the picture
- a **direction**, which refers to which way the arrow points.

In other words, the geometric picture of a vector always carries two pieces of information. (Here, I'm imagining we've drawn the vector as an arrow with one endpoint at the origin and pointing some way.)

In a lot of geometry situations we only care about the direction, and we want to ignore the magnitude. When that happens, one convention is to just set the magnitude equal to 1:

Definition

A **unit vector** will be a vector that has magnitude 1.

Thus we use the concept of unit vector to capture direction. So in $\mathbb{R}^2$, $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ is thought of as "due east" and $\begin{pmatrix} -1 \\ 0 \end{pmatrix}$ is "due west", while $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ is "due north" and $\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$ is "northeast".

Definition

Given any vector $\mathbf{v}$ in $\mathbb{R}^n$ besides the zero vector, the **direction along** $\mathbf{v}$ is the unit vector

$$\frac{\mathbf{v}}{|\mathbf{v}|}$$

which is the unit vector that points the same way that $\mathbf{v}$ does.

We will avoid referring to the direction of the zero-vector $\mathbf{0}$, which doesn't make sense. (If you try to apply the formula here, you get division by 0, since $\mathbf{0}$ is the only vector with length 0.) If you really want, you could say it has *every* direction, but this is a convention.

Warning

Depending on what book you're following, more pedantic authors might write "the unit vector in the direction of $\mathbf{v}$" or even "the unit vector in the same direction as $\mathbf{v}$" rather than "direction along $\mathbf{v}$". This is too long to type, so I adopted the shorter phrasing. I think everyone will know what you mean.

Type signature

If $\mathbf{v}$ is a nonzero vector of length n, then the direction along $\mathbf{v}$ is also a vector of length n.

Example

Let's first do examples in $\mathbb{R}^2$ so we can drawn some pictures.

- The direction along the vectors $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $\begin{pmatrix} 5 \\ 0 \end{pmatrix}$ or $\begin{pmatrix} 1337 \\ 0 \end{pmatrix}$ are all $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$, thought of as *due east.*
- But the direction along the vectors $\begin{pmatrix} -1 \\ 0 \end{pmatrix}$ or $\begin{pmatrix} -9 \\ 0 \end{pmatrix}$ are both $\begin{pmatrix} -1 \\ 0 \end{pmatrix}$, thought of as *due west.*
- The direction along the vectors $\begin{pmatrix} 0 \\ -2 \end{pmatrix}$, $\begin{pmatrix} 0 \\ -17 \end{pmatrix}$ are all $\begin{pmatrix} 0 \\ -1 \end{pmatrix}$, thought of as *due south.*

Example

How about the direction along $\begin{pmatrix} 3 \\ -4 \end{pmatrix}$? We need to first find the length of the vector so we can scale it down. That's given by the Pythagorean theorem, of course:

$$\left| \begin{pmatrix} 3 \\ -4 \end{pmatrix} \right| = \sqrt{3^2 + 4^2} = 5.$$

So the direction along $\begin{pmatrix} 3 \\ -4 \end{pmatrix}$ would be

$$\frac{1}{5} \begin{pmatrix} 3 \\ -4 \end{pmatrix} = \begin{pmatrix} \frac{3}{5} \\ -\frac{4}{5} \end{pmatrix}.$$

See Figure 1. The direction is somewhere between south and southeast.

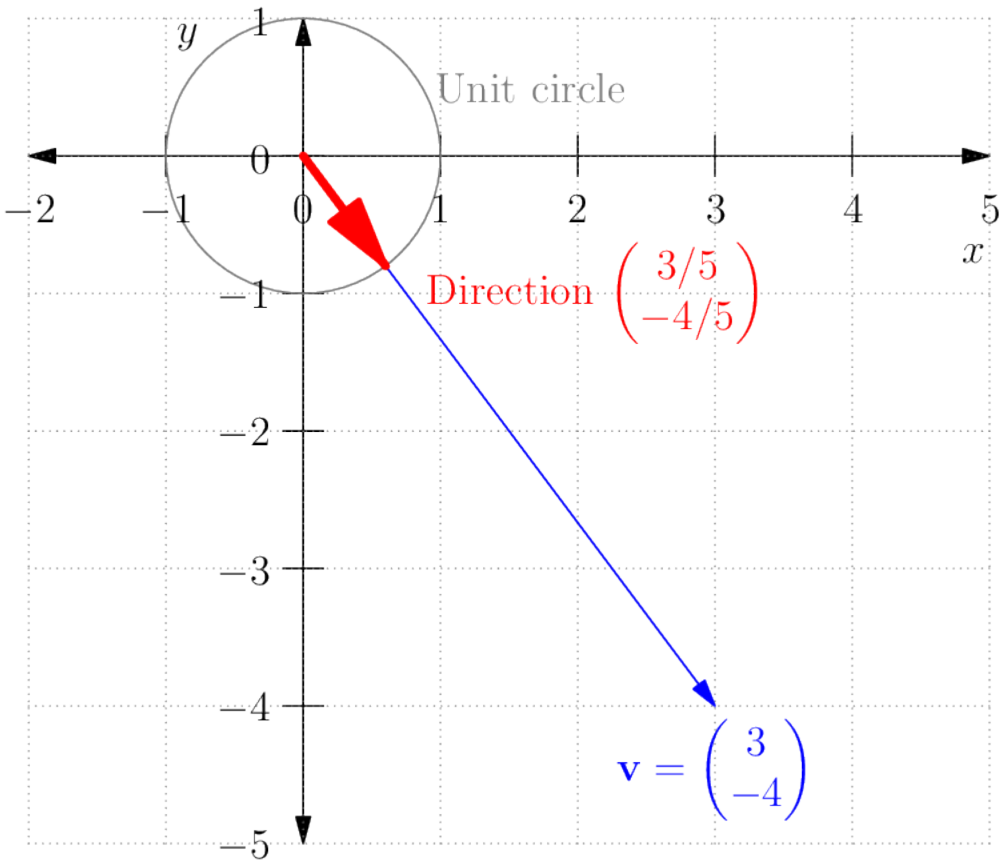

Figure 1: The direction along $\begin{pmatrix} 3 \\ -4 \end{pmatrix}$ (blue) is $\begin{pmatrix} 3/5 \\ -4/5 \end{pmatrix}$ (red). Unit vectors always lie on the grey circle (unit circle) by definition.

When drawn like Figure 1 in the two-dimensional picture $\mathbb{R}^2$, the notion of direction along $\begin{pmatrix} x \\ y \end{pmatrix}$ is *almost* like the notion of slope $\frac{y}{x}$ in high school algebra (so the slope of the blue ray in Figure 1). But there are a few reasons our notion of direction is more versatile than just using the slope of the blue ray.

- The notion of direction can tell the difference between $\begin{pmatrix} 3 \\ -4 \end{pmatrix}$, which goes southeast, and $\begin{pmatrix} -3 \\ 4 \end{pmatrix}$, which goes northwest. Slope cannot; it would assign both of these "slope $-\frac{4}{3}$.

- The due-north and due-south directions $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 0 \\ -1 \end{pmatrix}$ would have undefined slope due to division-by-zero, so you always have to worry about this extra edge case. With unit vectors, due-north and due-south don't cause extra headache.
- Our definition of direction works in higher dimension $\mathbb{R}^3$. There isn't a good analog of slope there; a single number cannot usefully capture a notion of direction in $\mathbb{R}^n$ for $n \geq 3$.

Example

The direction along the three-dimensional vector $\begin{pmatrix} 12 \\ -16 \\ 21 \end{pmatrix}$ is

$$\begin{pmatrix} 12/29 \\ -16/29 \\ 21/29 \end{pmatrix}.$$

See if you can figure out where the 29 came from.

§3.4 [RECIPE] Areas and volumes

If $\mathbf{v}_1 = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}$ and $\mathbf{v}_2 = \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}$ are vectors, drawn as arrows with a common starting point, then their sum $\mathbf{v}_1 + \mathbf{v}_2$ makes a parallelogram in the plane with $\mathbf{0}$ as shown in Figure 2.

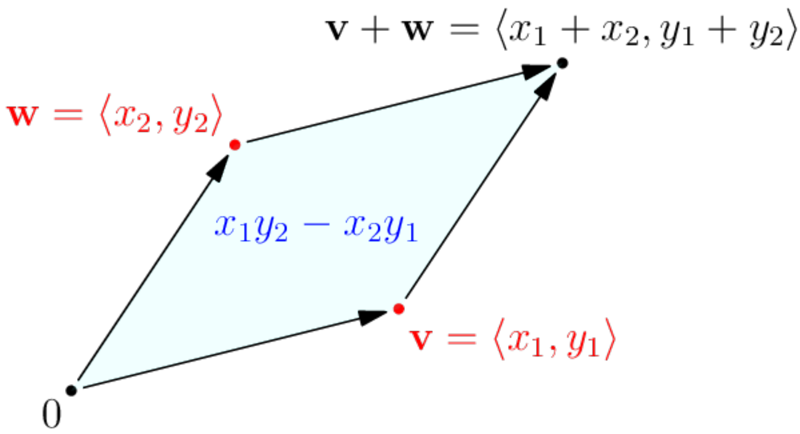

Figure 2: Vector addition in $\mathbb{R}^2$.

The following theorem is true, but we won't be able to prove it in 18.02.

Recipe for area of a parallelogram

The area of the parallelogram formed by $\mathbf{v}_1 = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}$ and $\mathbf{v}_2 = \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}$ is equal to the absolute value of the determinant

$$\det \begin{pmatrix} x_1 & x_2 \\ y_1 & y_2 \end{pmatrix} = x_1 y_2 - x_2 y_1.$$

A similar theorem is true for the parallelepiped[5] with three vectors in $\mathbb{R}^3$; see Figure 3.

[5] I hate trying to spell this word.

Recipe for volume of a parallelepiped

The volume of the parallelepiped formed by $\mathbf{v}_1 = \begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix}$, $\mathbf{v}_2 = \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix}$, $\mathbf{v}_3 = \begin{pmatrix} x_3 \\ y_3 \\ z_3 \end{pmatrix}$ is equal to the absolute value of the determinant

$$\det \begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{pmatrix}.$$

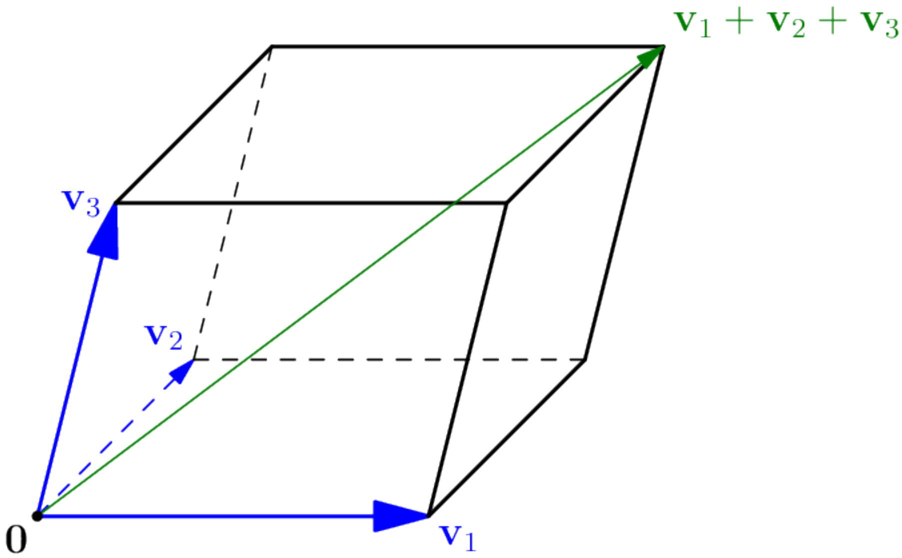

Figure 3: Three vectors in $\mathbb{R}^3$ making a parallelepiped.

Digression

If you're interested in the proof of these results and their n-dimensional generalizations, the tool needed is the **wedge product**, which is denoted $\bigwedge^{k(\mathbb{R}^n)}$. This is well beyond the scope of 18.02, but it's documented in Chapter 12 of my Napkin for those of you that want to read about it.

Alternatively, I think Wikipedia and Axler[6], among others, use a definition of the determinant as the unique multilinear alternating map on n-tuples of column vectors in $\mathbb{R}^n$ that equals 1 for the identity. This definition will work, and will let you derive the formula for determinant, and gives you a reason to believe it should match your concept of area and volume. It's probably also easier to understand than wedge products. However, in the long term I think wedge products are more versatile, even though they take much longer to setup.

§3.5 [EXER] Exercises

[6]Who has a paper called Down with Determinants!, that I approve of.

Exercise 3.1: Calculate the unit vector along the direction of the

$$\begin{pmatrix} -0.0008\pi \\ -0.0009\pi \\ -0.0012\pi \end{pmatrix}.$$

§4 The dot product

The dot product is the first surprising result you'll see in this class, because it has *two* definitions that look nothing alike, one algebraic and one geometric. Because of that, we'll be able to get a ton of mileage out of it.

This will be a general theme across the course: almost every new concept we define will have some sort "algebraic" side (like the coordinates for vector addition) and some "geometric" side (the parallelogram in Figure 2). This is the bar a concept has to pass for us to study it in this class: in order for us to deem a concept worthy of our attention in 18.02, it must have both an interpretation with algebra and an interpretation in geometry.

§4.1 [TEXT] Two different definitions of the dot product

I promised you two definitions right? So here they are.

Definition

Suppose $\mathbf{v} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$ and $\mathbf{w} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$ are two vectors in $\mathbb{R}^n$.

The *algebraic definition* is to take the sum of the component-wise products:

$$\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} := a_1 b_1 + \ldots + a_n b_n.$$

The *geometric definition* is that if θ is the angle between the two vectors when we draw them as arrows with a common starting point, then

$$\mathbf{v} \cdot \mathbf{w} := |\mathbf{v}| \, |\mathbf{w}| \cos\theta.$$

That is, the dot product equals the product of the lengths times the cosine of the included angle.

It's totally not obvious that these two definitions are the same? The standard proof uses the law of cosines; I'll say a bit more about this once I've done a few examples. I also found a proof without trigonometry that I typed in Section 27.2. I won't dwell on this proof too much in the interest of moving these notes forward.

Type signature

Remember, the dot product takes two vectors *of equal dimensions* as inputs and outputs a *scalar* (i.e. a real number). **It does not output a vector!** This is the mistake every calculus or linear algebra instructor dreads for the first few weeks of class.

Repeat: dot product output type is **number**! Not a vector!

Example

Let's find the dot product of $\mathbf{v} = \begin{pmatrix} -5 \\ 5\sqrt{3} \end{pmatrix}$ and $\mathbf{w} = \begin{pmatrix} 7\sqrt{3} \\ -7 \end{pmatrix}$, both ways.

- The algebraic definition is easy:

$$\mathbf{v} \cdot \mathbf{w} = -5 \cdot 7\sqrt{3} + 5\sqrt{3} \cdot (-7) = -70\sqrt{3}.$$

- The geometric definition is a bit more work, see Figure 4. In this picture, you can see there are two $30°$ angles between the axes, and the lengths of the vectors are 10 and 14. Hence, the angle θ between them is $\theta = 90° + (30° + 30°) = 150°$. So the geometric definition gives that

$$\mathbf{v} \cdot \mathbf{w} = |\mathbf{v}|\ |\mathbf{w}| \cos\theta = 10 \cdot 14 \cdot \cos(150°) = 140 \cdot -\frac{\sqrt{3}}{2} = -70\sqrt{3}.$$

TODO: an example of perpendicular in 2D

This example shows something new:

Memorize

Two nonzero vectors have perpendicular directions if and only if their dot product is 0.

TODO: an example of perpendicular in 3D

TODO: an example of lengths

Memorize

The dot product of a vector with itself is the squared length.

Tip

You can see from this example that computing the dot product of two given vectors with coordinates is way easier to do with the algebraic definition. This will be true in general throughout this class:

- Use the algebraic definition when you need to do practical calculation.
- Use the geometric definition to interpret the result in some way.

TODO

Figure 4: Some pictures of dot product.

§4.2 [SIDENOTE] The proof of the equivalence of the dot product properties

§4.3 [RECIPE] Checking whether two vectors are perpendicular

§4.4 [TEXT] Projection

§4.5 [RECIPE] Projection of one vector along the direction along another

Recipe for projecting one vector along another

Suppose $\mathbf{v}$ and $\mathbf{w}$ are given vectors in $\mathbb{R}^n$. To find the length of the projection of $\mathbf{v}$ along $\mathbf{w}$:

1. Output the absolute value of $\frac{\mathbf{v}\cdot\mathbf{w}}{|\mathbf{w}|}$.

To find the actual vector $\mathbf{v}$ along $\mathbf{w}$:

1. Output $\frac{\mathbf{v}\cdot\mathbf{w}}{|\mathbf{w}|}\frac{\mathbf{w}}{|\mathbf{w}|}$.

TODO: define it

§5 Planes and their normal vectors

§5.1 [TEXT] Planes in $\mathbb{R}^3$

§5.2 [TEXT] Normal vectors to lines in $\mathbb{R}^2$

Before we get to normal vectors to planes in $\mathbb{R}^3$, I want to do everything in $\mathbb{R}^2$ first.

If you are confused about normal vectors in the plane, it might help to first do the $\mathbb{R}^2$ case, which is easier to draw and for which you might have better intuition from eighth or ninth grade algebra.

Here's a question: which vectors in $\mathbb{R}^2$ are perpendicular to $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$? They're the vectors lying on a line of slope $-\frac{1}{2}$ through the origin, namely

$$0 = \begin{pmatrix} x \\ y \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 2 \end{pmatrix} \iff 0 = x + 2y.$$

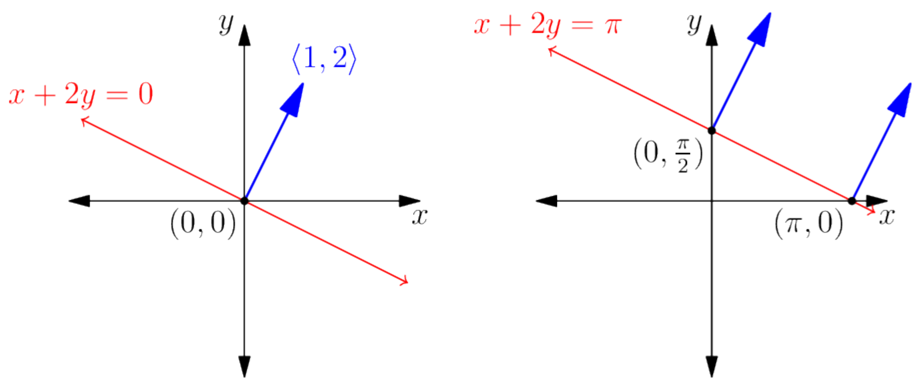

Figure 5: Plots of $x + 2y = 0$ and $x + 2y = \pi$.

Okay, in that case what does the line

$$x + 2y = \pi$$

look like? Well, it's a parallel line, the slope is still the same.

Equivalently, you could also imagine it as the points $\begin{pmatrix} x \\ y \end{pmatrix}$ such that

$$\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} \pi \\ 0 \end{pmatrix} \text{ is perpendicular to } (1, 2)$$

or do the same thing for any point on the line, like

$$\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} 0 \\ \pi/2 \end{pmatrix} \text{ is perpendicular to } (1, 2)$$

or even

$$\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} 0.218\pi \\ 0.564\pi \end{pmatrix} \text{ is perpendicular to } (1, 2)$$

But that's silly. Most of the time you don't care about base points. All you care is the line has slope $-\frac{1}{2}$, and for that the LHS just needs to be $x + 2y$ (or even $100x + 200y$). The RHS can be whatever you want.

In $\mathbb{R}^3$, the exact same thing is true for the expression $ax + by + cz = d$. The only difference is that the word "slope" is banned (or at least needs a new type; it won't be a single number). Nevertheless, even if we can't talk about slope, we can still talk about parallel planes, and now the whole discussion carries over wholesale.

§5.3 [RECIPE] Normal vectors to a plane

Idea

Everything we used slope for in 18.01, we should rephrase in terms of normal vectors for 18.02.

Recipe for calculating normal vector to a plane

To find the normal vector of a plane given in the form $ax + by + cz = d$:

1. Output $\begin{pmatrix} a \\ b \\ c \end{pmatrix}$ (or any other multiple of $\begin{pmatrix} a \\ b \\ c \end{pmatrix}$).

§5.4 [RECIPE] Finding a plane through a point with a direction

Recipe for finding a plane given a normal vector and a point on it

Suppose the given normal vector is $\begin{pmatrix} a \\ b \\ c \end{pmatrix}$, and $P \in \mathbb{R}^3$ is a given point.

1. Write $ax + by + cz$ for the left-hand side.
2. Evaluate the left-hand side at P to get a number d.
3. Output $ax + by + cz = d$.

Sample Question

Find the equation of the plane parallel to $x + 2y + 3z = 100$ which passes through the point $(1, 4, 9)$.

Solution

Planes are parallel when they have the same normal vector, so we know the normal vector is $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ for both. Hence the answer should take the form $x + 2y + 3z = d$ for some d. In order to pass through $(1, 4, 9)$ we should choose $d = 1 + 2 \cdot 4 + 3 \cdot 9 = 36$. So output $x + 2y + 3z = 36$.

§5.5 [TEXT] Calculating distance to a plane

§5.6 [RECIPE] Distance to a plane

§6 The cross product

The cross product is the last major linear algebra tool we'll need to introduce (together with determinants and the dot product). Like the dot product, the cross product also has two definitions, one algebraic and one geometric.

§6.1 [TEXT] The two definitions of the cross product

§6.2 [SIDENOTE] The cross product sucks

Compared to dot products and determinants, the cross product might feel the most unnatural, for good reason — it's used much less frequently by serious mathematicians than the other tools you see.

Figure 6: How to think of cross products.

The reason that the cross product isn't popular with mathematicians is the definition of the cross product is **really quite brittle**. For example, the cross product can't be defined for any number of dimensions,[7] and you have to remember this weird right-hand rule that adds one more arbitrary convention. So the definition is pretty unsatisfying.

To replace the cross product, mathematicians use a different kind of object called a *bivector*, an element of a space called $\bigwedge^2(\mathbb{R}^n)$. (They might even claim that bivectors do everything cross products can do, but better.) Again, this new kind of object is well beyond the scope of 18.02 but it's documented in Chapter 12 of my Napkin if you do want to see it.

I'll give you a bit of a teaser though. In general, for any n, bivectors in $\mathbb{R}^n$ are specified by $\frac{n(n-1)}{2}$ coordinates. So for $n = 3$ you *could* translate every bivector in $\mathbb{R}^3$ into a vector in $\mathbb{R}^3$ by just reading the coordinates (although you end up with the right-hand rule as an artifact of the translation), and the cross product is exactly what you get. But for $n = 4$, a bivector in $\mathbb{R}^4$ has six numbers, which is

[7]Just kidding, apparently there's a seven dimensional cross product? Today I learned. Except that there are apparently 480 different ways to define it in seven dimensions, so, like, probably not a great thing.

too much information to store in a vector in $\mathbb{R}^4$. Similarly, for $n > 4$, this translation can't be done. That's why the cross product is so brittle and can't work past $\mathbb{R}^3$.

§6.3 [RECAP] Recap of vector stuff up to here

Part Bravo: Linear Algebra of Matrices

For comparison, this part corresponds approximately to §3, §4, §6 of Poonen's notes.

§7 Linear transformations and matrices

This section will be presented a bit differently than you'll see in many other places; I talk about linear transformations first, and then talk about matrices as an encoding of linear transformations. I feel quite strongly this way is better, but if you are in an actual course, their presentation is likely to be different (and worse).

§7.1 [TEXT] Linear transformation

The definition I'm about to give is the 18.700/18.701 definition of linear transform, but the hill I will die on is that this definition is better than the one 18.02.

Definition of linear transformation

A *linear transform* $T : \mathbb{R}^n \to \mathbb{R}^m$ is *any* map obeying the two axioms $T(c\mathbf{v}) = cT(\mathbf{v})$ and $T(\mathbf{v} + \mathbf{w}) = T(\mathbf{v}) + T(\mathbf{w})$.

So it's a chonky boy: for every $\mathbf{v} \in \mathbb{R}^n$, there's an output value $T(\mathbf{v}) \in \mathbb{R}^m$. I wouldn't worry too much about the axioms until later; for now, read the examples.

Examples of linear transformations

The following are all linear transformations from $\mathbb{R}^2$ to $\mathbb{R}^2$:

- The constant function where $T(\mathbf{v}) = \mathbf{0}$ for every vector v
- Projection onto the x-axis: $T\left(\binom{x}{y}\right) = \binom{x}{0}$.
- Rotation by an angle
- Reflection across a line
- Projection onto the line $y = x$.
- Multiplication by any 2×2 matrix, e.g. the formula

$$T\left(\binom{x}{y}\right) = \begin{pmatrix} x + 2y \\ 3x + 4y \end{pmatrix}$$

  is a linear transformation too.

Tip

Note that $T(\mathbf{0}) = \mathbf{0}$ in any linear transformation.

The important principle to understand is that if you know the values of a transformation T at enough points, you can recover the rest.

Here's an easy example to start:

Question 7.1: If $T : \mathbb{R}^2 \to \mathbb{R}^2$ is a linear transform and it's given that

$$T\left(\begin{pmatrix} 3 \\ 4 \end{pmatrix}\right) = \begin{pmatrix} \pi \\ 9 \end{pmatrix} \text{ and } T\left(\begin{pmatrix} 100 \\ 100 \end{pmatrix}\right) = \begin{pmatrix} 0 \\ 12 \end{pmatrix}$$

what are the vectors for $T\left(\begin{pmatrix} 103 \\ 104 \end{pmatrix}\right)$ and $T\left(\begin{pmatrix} 203 \\ 204 \end{pmatrix}\right)$?

Solution:

$$T\left(\begin{pmatrix} 103 \\ 104 \end{pmatrix}\right) = \begin{pmatrix} \pi \\ 9 \end{pmatrix} + \begin{pmatrix} 0 \\ 12 \end{pmatrix} = \begin{pmatrix} \pi \\ 21 \end{pmatrix}$$
$$T\left(\begin{pmatrix} 203 \\ 204 \end{pmatrix}\right) = \begin{pmatrix} \pi \\ 9 \end{pmatrix} + 2\begin{pmatrix} 0 \\ 12 \end{pmatrix} = \begin{pmatrix} \pi \\ 33 \end{pmatrix}.$$

□

Here's another example.

Question 7.2: If $T : \mathbb{R}^2 \to \mathbb{R}^2$ is a linear transform and it's given that

$$T\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}\right) = \begin{pmatrix} 1 \\ 3 \end{pmatrix} \text{ and } T\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$$

what is $T\left(\begin{pmatrix} 50 \\ 70 \end{pmatrix}\right)$?

Solution:

$$T\left(\begin{pmatrix} 50 \\ 70 \end{pmatrix}\right) = 50\begin{pmatrix} 1 \\ 3 \end{pmatrix} + 70\begin{pmatrix} 2 \\ 4 \end{pmatrix} = \begin{pmatrix} 190 \\ 430 \end{pmatrix}.$$

□

More generally, the second question shows that if you know $T\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}\right)$ and $T\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right)$ you ought to be able to *calculate* the output of T at any other vector like $\begin{pmatrix} 50 \\ 70 \end{pmatrix}$. To expand on this:

$$T\left(\begin{pmatrix} a \\ b \end{pmatrix}\right) = aT\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}\right) + bT\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right). \quad (2)$$

More generally, from understanding the solution to the above two questions, you should understand the following important statement that we'll use over and over.

! Memorize

Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation. If you know the outputs T on a basis, then you can deduce the value of T at any other input.

For now "basis" refers to just the n vectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$. But later on we will generalize this notion to some other settings too.

§7.2 [RECIPE] Matrix encoding

A *matrix* is a way of *encoding* the *outputs* of T using as few numbers as possible. That is:

Definition

A matrix **encodes all outputs** of a linear transformation T by **writing the outputs** of $T(\mathbf{e}_1)$, ..., $T(\mathbf{e}_n)$ as a list of **column vectors**.

For example, if you had $T : \mathbb{R}^2 \to \mathbb{R}^2$ with

$$T\left(\begin{pmatrix}1\\0\end{pmatrix}\right) = \begin{pmatrix}1\\3\end{pmatrix} \quad \text{and } T\left(\begin{pmatrix}0\\1\end{pmatrix}\right) = \begin{pmatrix}2\\4\end{pmatrix} \iff T \quad \text{encoded as} \quad \begin{pmatrix}1 & 2\\3 & 4\end{pmatrix}.$$

To put this into recipe form:

Definition for encoding a transformation

Given a transformation $T : \mathbb{R}^n \to \mathbb{R}^m$, to encode it as a matrix:

1. Compute $T(\mathbf{e}_1)$ through $T(\mathbf{e}_n)$ and write them as column vectors..
2. Glue them together to get an $n \times m$ array of numbers.

Here's more examples.

Sample Question

Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be projection onto the x-axis. Write T as a 2×2 matrix.

Solution: Note that

$$T\left(\begin{pmatrix}1\\0\end{pmatrix}\right) = \begin{pmatrix}1\\0\end{pmatrix} \qquad T\left(\begin{pmatrix}0\\1\end{pmatrix}\right) = \begin{pmatrix}0\\0\end{pmatrix}.$$

Glue these together and output T as the matrix

$$T = \begin{pmatrix}1 & 0\\0 & 0\end{pmatrix}.$$

□

Remark

You might note that indeed multiplication by the encoded matrix

$$\begin{pmatrix}1 & 0\\0 & 0\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}x\\0\end{pmatrix}.$$

matches what you expect: $\left(\begin{smallmatrix}x\\0\end{smallmatrix}\right)$ is indeed the projection of $\left(\begin{smallmatrix}x\\y\end{smallmatrix}\right)$ onto the x-axis! And this works for every linear transformation. This is so important I'll say it again next section, just mentioning it here first.

Sample Question

Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be reflection around the line $y = x$. Write T as a 2×2 matrix.

Solution: Note that

$$T\left(\begin{pmatrix}1\\0\end{pmatrix}\right)=\begin{pmatrix}0\\1\end{pmatrix} \qquad T\left(\begin{pmatrix}0\\1\end{pmatrix}\right)=\begin{pmatrix}1\\0\end{pmatrix}.$$

Glue these together and output T as the matrix

$$T=\begin{pmatrix}0&1\\1&0\end{pmatrix}.$$

□

Sample Question

Let $T:\mathbb{R}^2\to\mathbb{R}^2$ be counterclockwise rotation around the origin by 30°. Write T as a 2×2 matrix.

Solution: See Figure 7. By looking at the unit circle, we see that

$$T(\mathbf{e_1})=\begin{pmatrix}\cos 30^\circ\\ \sin 30^\circ\end{pmatrix}=\left(\frac{\sqrt{3}}{2},\frac{1}{2}\right).$$

The vector $\mathbf{e}_2$ is 90° further along

$$T(\mathbf{e_2})=\begin{pmatrix}\cos 120^\circ\\ \sin 120^\circ\end{pmatrix}=\left(-\frac{1}{2},\frac{\sqrt{3}}{2}\right).$$

Glue these together and output T as the matrix

$$T=\begin{pmatrix}\frac{\sqrt{3}}{2}&-\frac{1}{2}\\ \frac{1}{2}&\frac{\sqrt{3}}{2}\end{pmatrix}.$$

□

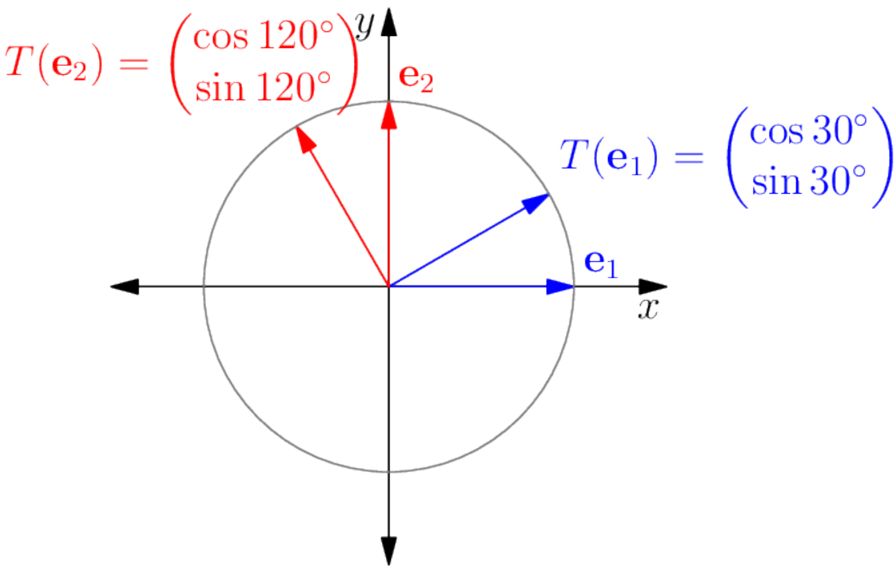

Figure 7: Rotation by 30 degrees.

Remark: This is where the rotation matrix comes from

If you redo this question with 30° replaced by any angle θ, you get the answer

$$T = \begin{pmatrix} \cos(\theta) & \cos(\theta + 90^\circ) \\ \sin(\theta) & \sin(\theta + 90^\circ) \end{pmatrix}.$$

So this is the matrix that corresponds to rotation. However, in the literature you will often see this rewritten as

$$T = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$$

to get rid of the $+90^\circ$ offsets. That's fine, but I think it kind of hides where the formula for rotation matrix comes from, personally.

Another example is the identity function:

Example: The identity matrix deserves its name

Let $I : \mathbb{R}^3 \to \mathbb{R}^3$ denote the 3D identity function, meaning $I(\mathbf{v}) = \mathbf{v}$. To encode it, we look at its values at $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$:

$$I(\mathbf{e}_1) = \mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad I(\mathbf{e}_2) = \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad I(\mathbf{e}_3) = \mathbf{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

We encode it as a matrix by writing the columns side by side, getting what you expect:

$$I \quad \text{encoded as} \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

This gives a more natural reason why the identity matrix is the one with 1's on the diagonal and 0's elsewhere (compared to the "well try multiplying by it" you learned in high school).

§7.3 [TEXT] Matrix multiplication

In the prerequisites, I said that you were supposed to know the rule for multiplying matrices, so you should already know for example that

$$\begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 23 & 34 \\ 31 & 46 \end{pmatrix}.$$

The goal of this section is to now explain why matrix multiplication is defined in this funny way. We will see two results:

- Multiplication of the matrix for T by a column vector $\mathbf{v}$ corresponds to evaluation $T(\mathbf{v})$.
- Multiplication of the matrices for S and T gives the matrix for the composed function $S \circ T$.[8]

[8] The $\circ$ symbol means the function where you apply T first then S first. So for example, if $f(x) = x^2$ and $g(x) = x + 5$, then $(f \circ g)(x) = f(g(x)) = (x+5)^2$. We mostly use that circle symbol if we want to refer to $f \circ g$ itself without the x, since it would look really bad if you wrote "$f(g$" or something.

§7.3.1 One matrix

Recall from Question 7.2 that if T was the linear transformation for which

$$T(\mathbf{e}_1) = \begin{pmatrix} 1 \\ 3 \end{pmatrix} \quad \text{and } T(\mathbf{e}_2) = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$$

then

$$T\left(\begin{pmatrix} 50 \\ 70 \end{pmatrix}\right) = \begin{pmatrix} 190 \\ 430 \end{pmatrix}.$$

We just now also saw that to encode T as a matrix, we have

$$T = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}.$$

Now, what do you think happens if you compute

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 50 \\ 70 \end{pmatrix}$$

as you were taught in high school? Surprise: you get $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}\begin{pmatrix} 50 \\ 70 \end{pmatrix} = \begin{pmatrix} 1 \cdot 50 + 2 \cdot 70 \\ 3 \cdot 50 + 4 \cdot 70 \end{pmatrix} = \begin{pmatrix} 190 \\ 340 \end{pmatrix}$ which is not just the same answer, but also the same intermediate calculations. In other words,

> **Idea**
>
> If one multiplies a matrix M by a column vector $\mathbf{v}$, this corresponds to applying the linear transformation T encoded by M to $\mathbf{v}$.

§7.3.2 Two matrices

Now, any time we have functions in math, we can *compose* them. So let's play the same game with a pair of functions S and T, and think about their composition $S \circ T$. Imagine we got asked the following question:

Question 7.3: Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be a linear transform such that

$$T\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}\right) = \begin{pmatrix} 1 \\ 3 \end{pmatrix} \quad \text{and } T\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} 2 \\ 4 \end{pmatrix}.$$

Then let $S : \mathbb{R}^2 \to \mathbb{R}^2$ be the linear transform such that

$$S\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}\right) = \begin{pmatrix} 5 \\ 7 \end{pmatrix} \quad \text{and } S\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} 6 \\ 8 \end{pmatrix}.$$

Evaluate $S\left(T\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}\right)\right)$ and $S\left(T\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right)\right)$.

Solution:

$$\begin{aligned} S\left(T\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}\right)\right) &= S\left(\begin{pmatrix} 1 \\ 3 \end{pmatrix}\right) = 1\begin{pmatrix} 5 \\ 7 \end{pmatrix} + 3\begin{pmatrix} 6 \\ 8 \end{pmatrix} = \begin{pmatrix} 23 \\ 31 \end{pmatrix} \\ S\left(T\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right)\right) &= S\left(\begin{pmatrix} 2 \\ 4 \end{pmatrix}\right) = 2\begin{pmatrix} 5 \\ 7 \end{pmatrix} + 4\begin{pmatrix} 6 \\ 8 \end{pmatrix} = \begin{pmatrix} 34 \\ 46 \end{pmatrix}. \end{aligned}$$

□

Now, $S \circ T$ is *itself* a function, so it makes sense to encode $S \circ T$ as a matrix too, using the answer to Question 7.3:

$$S\left(T\left(\begin{pmatrix}1\\0\end{pmatrix}\right)\right) = \begin{pmatrix}23\\31\end{pmatrix} \text{ and } S\left(T\left(\begin{pmatrix}0\\1\end{pmatrix}\right)\right) = \begin{pmatrix}34\\46\end{pmatrix} \Longleftrightarrow S \circ T \quad \text{encoded as} \quad \begin{pmatrix}23 & 34\\31 & 46\end{pmatrix}.$$

The matrix multiplication rule is then rigged to give the same answer through the same calculation again:

$$\begin{pmatrix}5 & 6\\7 & 8\end{pmatrix}\begin{pmatrix}1 & 2\\3 & 4\end{pmatrix} = \begin{pmatrix}23 & 34\\31 & 46\end{pmatrix}.$$

In other words:

Idea

If one multiplies two matrices M and N, this corresponds to composing the linear transformations that M and N encode.

This shows why the 18.700/18.701 definitions are better than the 18.02 ones. In 18.02, the recipe for matrix multiplication is a *definition*: "here is this contrived rule about taking products of columns and rows, trust me bro". But in 18.700/18.701, the matrix multiplication recipe is a *theorem*; it's what happens if you generalize Question 7.3 to eight variables (or $n^2 + n^2 = 2n^2$ variables for $n \times n$ matrices).

Digression

As an aside, this should explain why matrix multiplication is associative but not commutative:

- Because function composition is associative, so is matrix multiplication.
- Because function composition is *not* commutative in general, matrix multiplication isn't either.

§7.4 [EXER] Exercises

Exercise 7.4: If A is a 3×3 matrix with determinant 2, what values could $\det(10A)$ take?

§8 Linear combinations of vectors

§8.1 [TEXT] Basis of vectors

§8.2 [RECIPE] Describing spans of explicit vectors

§8.3 Systems of equations

§8.4 [RECIPE] Number of solutions to a square system of linear equations

§9 Eigenvalues and eigenvectors

§9.1 [TEXT] The problem of finding eigenvectors

Let's define the relevant term first:

Definition

Suppose T is a matrix or linear transformation, λ a scalar, and $\mathbf{v}$ is a vector such that

$$T(\mathbf{v}) = \lambda \mathbf{v};$$

that is, T sends $\mathbf{v}$ to a multiple of itself. Then we call λ an **eigenvalue** and $\mathbf{v}$ an **eigenvector**.

Type signature

Eigenvalues λ are always scalars.

Example

Let $T = \begin{pmatrix} 74 & 52 \\ 32 & 36 \end{pmatrix}$ and consider the vector $\mathbf{v} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$. Then

$$T(\mathbf{v}) = \begin{pmatrix} 74 & 52 \\ 32 & 36 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 200 \\ 100 \end{pmatrix} = 100 \begin{pmatrix} 2 \\ 1 \end{pmatrix} = 100\mathbf{v}.$$

So we would say $\mathbf{v}$ is an eigenvector with eigenvalue 100.

Of course, if $\mathbf{v}$ is an eigenvector, so are all its multiples, e.g.

$$\begin{pmatrix} 74 & 52 \\ 32 & 36 \end{pmatrix} \begin{pmatrix} 20 \\ 10 \end{pmatrix} = \begin{pmatrix} 2000 \\ 1000 \end{pmatrix} = 100 \begin{pmatrix} 20 \\ 10 \end{pmatrix}$$

so $\begin{pmatrix} 20 \\ 10 \end{pmatrix}$ is an eigenvector with the same eigenvalue 100, etc.

Remark

The stupid solution $\mathbf{v} = \mathbf{0}$ always satisfies the eigenvector equation for any λ, so we will pretty much ignore it and focus only on finding nonzero eigenvectors.

The goal of this section is to show:

Question

Given an encoding of T as a matrix, how can we find its eigenvectors (besides $\mathbf{0}$)?

§9.2 [TEXT] How to come up with the recipe for eigenvalues

For this story, our protagonist will be the matrix

$$A = \begin{pmatrix} 5 & -2 \\ 3 & 10 \end{pmatrix}.$$

Phrased another way, the problem of finding eigenvectors is, by definition, looking for λ, x, y such that

$$A\begin{pmatrix} x \\ y \end{pmatrix} = \lambda\begin{pmatrix} x \\ y \end{pmatrix} \iff \begin{cases} 5x - 2y = \lambda x \\ 3x + 10y = \lambda y \end{cases}.$$

Smart-alecks will say $x = y = 0$ always works for every λ. *Are there other solutions?*

§9.2.1 Why guessing the eigenvalues is ill-fated

As an example, let's see if there are any eigenvectors $\binom{x}{y}$ with eigenvalue 100. In other words, let's solve

$$\begin{pmatrix} 5 & -2 \\ 3 & 10 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = 100\begin{pmatrix} x \\ y \end{pmatrix}.$$

If we solve the system of equations, we get

$$\begin{cases} 5x - 2y = 100x \\ 3x + 10y = 100y \end{cases} \implies \begin{cases} -95x - 2y = 0 \\ 3x - 90y = 0 \end{cases} \implies x = y = 0.$$

Well, that's boring. In this system of equations, the only solution is $x = y = 0$.

We can try a different guess: maybe we use 1000 instead of 100. An eigenvector with eigenvalue 1000 ought to be correspond to

$$\begin{pmatrix} 5 & -2 \\ 3 & 10 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = 1000\begin{pmatrix} x \\ y \end{pmatrix}.$$

If we solve the system of equations, we get

$$\begin{cases} 5x - 2y = 1000x \\ 3x + 10y = 1000y \end{cases} \implies \begin{cases} -995x - 2y = 0 \\ 3x - 990y = 0 \end{cases} \implies x = y = 0$$

which… isn't any better. We still don't get any solutions besides $x = y = 0$.

At this point, you should be remembering something I told you in R04: a "random" system of equations and variables usually only has a unique solution. So if I keep picking numbers out of a hat like 100, 1000, etc., then I'm unlikely to find anything interesting. In order to get a system that doesn't just solve to $x = y = 0$, I'm going to need to cherry-pick my number λ.

§9.2.2 Cherry-picking λ

Let's try to figure out what value of λ would make the system more interesting. If we copy what we did above, we see that the general process is:

$$\begin{cases} 5x - 2y = \lambda x \\ 3x + 10y = \lambda y \end{cases} \implies \begin{cases} (5-\lambda)x - 2y = 0 \\ 3x + (10-\lambda)y = 0 \end{cases}$$

We need to cherry-pick λ to make sure that the system doesn't just solve to $x = y = 0$ like the examples we tried with 100 and 1000. But we learned how to do this in R04: in order to get a degenerate system you need to make sure that

$$0 = \det\begin{pmatrix} 5-\lambda & -2 \\ 3 & 10-\lambda \end{pmatrix}.$$

> **Remark**
>
> At this point, you might notice that this is secretly an explanation of why $A - \lambda I$ keeps showing up on your formula sheet. Writing $A\mathbf{v} = \lambda\mathbf{v}$ is the same as $(A - \lambda I)\mathbf{v} = 0$, just more opaquely.

Expanding the determinant on the left-hand side gives

$$0 = \det\begin{pmatrix} 5-\lambda & -2 \\ 3 & 10-\lambda \end{pmatrix} = (5-\lambda)(10-\lambda) + 6 = \lambda^2 - 15\lambda + 56 = (\lambda - 7)(\lambda - 8).$$

Great! So we expect that if we choose either $\lambda = 7$ and $\lambda = 8$, then we will get a degenerate system, and we won't just get $x = y = 0$. Indeed, let's check this:

- When $\lambda = 7$, our system is

  $$\begin{cases} 5x - 2y = 7x \\ 3x + 10y = 7y \end{cases} \implies \begin{cases} -2x - 2y = 0 \\ 3x + 3y = 0 \end{cases} \implies x = -y.$$

  So for example, $\left(\begin{smallmatrix} -13 \\ 13 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 37 \\ -37 \end{smallmatrix}\right)$ will be eigenvectors with eigenvalue 7:

  $$A\begin{pmatrix} -13 \\ 13 \end{pmatrix} = \begin{pmatrix} 5 & -2 \\ 3 & 10 \end{pmatrix}\begin{pmatrix} -13 \\ 13 \end{pmatrix} = \begin{pmatrix} -91 \\ 91 \end{pmatrix} = 7\begin{pmatrix} -13 \\ 13 \end{pmatrix}.$$

  On exam, you probably answer "the eigenvectors with eigenvalue 7 are the multiples of $\left(\begin{smallmatrix} -1 \\ 1 \end{smallmatrix}\right)$", or "the eigenvectors with eigenvalue 7 are the multiples of $\left(\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right)$" if you want; these are the same thing. Or if you want to mess with the grader, "the eigenvectors with eigenvalue 7 are the multiples of $\left(\begin{smallmatrix} 100 \\ -100 \end{smallmatrix}\right)$" is fine too.

- When $\lambda = 8$, our system is

  $$\begin{cases} 5x - 2y = 8x \\ 3x + 10y = 8y \end{cases} \implies \begin{cases} -3x - 2y = 0 \\ 3x + 2y = 0 \end{cases} \implies x = -\frac{2}{3}y.$$

  So for example, $\left(\begin{smallmatrix} -20 \\ 30 \end{smallmatrix}\right)$ is an eigenvector with eigenvalue 8:

  $$A\begin{pmatrix} -20 \\ 30 \end{pmatrix} = \begin{pmatrix} 5 & -2 \\ 3 & 10 \end{pmatrix}\begin{pmatrix} -20 \\ 30 \end{pmatrix} = \begin{pmatrix} -160 \\ 240 \end{pmatrix} = 8\begin{pmatrix} -20 \\ 30 \end{pmatrix}.$$

  On exam, you should answer "the eigenvectors with eigenvalue 8 are the multiples of $\left(\begin{smallmatrix} -2 \\ 3 \end{smallmatrix}\right)$". Or you can say "the eigenvectors with eigenvalue 8 are the multiples of $\left(\begin{smallmatrix} 2 \\ -3 \end{smallmatrix}\right)$" if you want; these are the same thing. You could even say "the eigenvectors with eigenvalue 8 are the multiples of $\left(\begin{smallmatrix} 200 \\ -300 \end{smallmatrix}\right)$" and still get credit, but that's silly.

§9.3 [RECAP] Summary

To summarize the story above:

- We had the matrix $A = \left(\begin{smallmatrix} 5 & -2 \\ 3 & 10 \end{smallmatrix}\right)$ and wanted to find λ's for which the equation

  $$\begin{pmatrix} 5 & -2 \\ 3 & 10 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \lambda\begin{pmatrix} x \\ y \end{pmatrix}$$

  had solutions other than $x = y = 0$.

- We realized that guessing λ was never going to fly, so we went out of our way to cherry-pick λ to make sure the system was degenerate. The buzzwords for this are "find the roots of the characteristic polynomial", but I wanted to show that it flows naturally from the end goal.
- For the two values of λ we cherry-picked, we know the system of equations is degenerate. So we solve the two degenerate systems and see what happens.

In lectures and notes, the last two bullets are separated as two different steps, to make it into a recipe. But don't lose sight of how they're connected! I would rather call it the following interlocked thing:

We cherry-pick λ to make sure the system doesn't just solve to $x = y = 0$.

To do the cherry-picking, ensure $\det(A - \lambda I) = 0$ so that our system is degenerate.

§9.4 [RECIPE] Calculating all the eigenvalues

To repeat the story:

Recipe for finding the eigenvectors and eigenvalues

Given a matrix A, to find its eigenvectors and eigenvalues:

1. Find all the values of λ such that, if you subtract λ from every diagonal entry of A (that is, look at $A - \lambda I$), the resulting square matrix of coefficients has determinant 0.
2. For each λ, solve the degenerate system and output the solutions to it. (You should find there is at least a one-dimensional space of solutions.)

TODO: Write example

§9.5 [TEXT] Solving degenerate systems

When carrying out the recipe for finding eigenvectors and eigenvalues, after cherry-picking λ, you have to solve a degenerate system of equations. Since most of the systems of equations you encounter in practice are nondegenerate, here's a few words of advice on instincts for solving the degenerate ones.

§9.5.1 Degenerate systems of two equations all look stupid

This is worth repeating: **degenerate systems of two equations all look stupid**. Earlier on, we saw the two systems

$$\begin{cases} -2x - 2y = 0 \\ 3x + 3y = 0 \end{cases} \qquad \text{and} \qquad \begin{cases} -3x - 2y = 0 \\ 3x + 2y = 0 \end{cases}.$$

Both look moronic to the eye, because in each equation, the two equations say the same thing. This is by design: when you're solving the eigenvector problem, *you're going out of your way to find degenerate systems* so that there will actually be solutions besides $x = y = 0$.

In particular: if you do all the steps right, **you should never wind up with $x = y = 0$ as your only solution**. That means you either didn't do the cherry-picking step correctly, or something went wrong when you were solving the system. If that happens, check your work!

§9.5.2 Degenerate systems of three equations may not look stupid, but they are

When you have three or more equations instead, they don't necessarily look as stupid. To reuse the example I mentioned from R04, we have

$$\begin{aligned} x + 10y - 9z &= 0 \\ 3x + y + 10z &= 0 \\ 4x + 11y + z &= 0 \end{aligned}$$

which doesn't look stupid. But again, if you check the determinant, you find out

$$\det \begin{pmatrix} 1 & 10 & -9 \\ 3 & 1 & 10 \\ 4 & 11 & 1 \end{pmatrix} = 0.$$

So you know *a priori* that there will be solutions besides $x = y = z = 0$.

I think 18.02 won't have too many situations where you need to solve a degenerate three-variable system of equations, because it's generally annoying to do by hand. But if it happens, you should fall back on your high school algebra and solve the system however you learned it in 9th or 10th grade. The good news is that at least one of the three equations is redundant, so you can just throw one away and solve for the other two. For example, in this case we would solve

$$\begin{aligned} x + 10y &= 9z \\ 3x + y &= -10z \end{aligned}$$

for x and y, as a function of z. I think this particular example works out to $x = -\frac{109}{29}z$, $y = \frac{37}{29}z$. And it indeed fits the third equation too.

§9.6 [SIDENOTE] Complex eigenvectors

Even in the 2×2 case, you'll find a lot of matrices M with real coefficients don't have eigenvectors. Here's one example.

Let

$$M = \begin{pmatrix} \cos(60°) & -\sin(60°) \\ \sin(60°) & \cos(60°) \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix}.$$

be the matrix corresponding to rotation by 60 degrees. (Feel free to replace 60 by a different number.) I claim that M has no real eigenvalues or eigenvectors.

Indeed, if $\mathbf{v} \in \mathbb{R}^2$ was an eigenvector, then $M\mathbf{v}$ needs to point in the same direction as $\mathbf{v}$, by definition. But that can never happen: M is rotation by 60°, so $M\mathbf{v}$ and $\mathbf{v}$ necessarily point in different directions — 60 degrees apart.

TODO: what goes wrong?

§9.7 [SIDENOTE] Application of eigenvectors: matrix powers

This is off-syllabus for 18.02, but I couldn't resist including it because it shows you a good use of eigenvalues in a seemingly unrelated problem, and also reinforces the idea that I keep axe-grinding:

If you have a linear operator T, and you know the outputs of T on *any* basis, that tells you all the outputs of T.

The problem is this:

? Question

Let M be the matrix $\begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}$. Calculate M^{100}.

At first glance, you might thinks question is obviously impossible without a computer, because raising a matrix to the 100th power would require 100 matrix multiplications. But I'll show you how to do it with eigenvectors.

Solution: First, we compute the eigenvectors and eigenvalues of M. If you follow the recipe, you'll get the following results:

- The vector $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ is an eigenvector with eigenvalue 2 (as is any multiple of $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$), because $M\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \end{pmatrix} = 2\begin{pmatrix} 1 \\ 0 \end{pmatrix}$.
- The vector $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ is an eigenvector with eigenvalue 3 (as is any multiple of $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$), because $M\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \end{pmatrix} = 3\begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

Now the trick is the following: it's really easy to apply M^{100} to the *eigenvectors*, because it's just multiplication by a constant. For example, the first few powers of M on $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ each double the vector, since they are all eigenvectors with eigenvalue 2; that is:

$$M\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$$
$$M^2\begin{pmatrix} 1 \\ 0 \end{pmatrix} = M\begin{pmatrix} 2 \\ 0 \end{pmatrix} = \begin{pmatrix} 4 \\ 0 \end{pmatrix}$$
$$M^3\begin{pmatrix} 1 \\ 0 \end{pmatrix} = M\begin{pmatrix} 4 \\ 0 \end{pmatrix} = \begin{pmatrix} 8 \\ 0 \end{pmatrix} \vdots$$

and so on, until

$$M^{100}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = 2^{100}\begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

By the same token:

$$M^{100}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = 3^{100}\begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

So now we know the outputs of M^{100} at two linearly independent vectors. It would be sufficient, then, to use this information to extract $M^{100}(\mathbf{e}_1)$ and $M^{100}(\mathbf{e}_2)$. We can now rewrite this as

$$M^{100}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2^{100} \\ 0 \end{pmatrix}; \qquad M^{100}\begin{pmatrix} 0 \\ 1 \end{pmatrix} = M^{100}\begin{pmatrix} 1 \\ 1 \end{pmatrix} - M^{100}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 3^{100} - 2^{100} \\ 3^{100} \end{pmatrix}.$$

Thus encoding M gives the answer:

$$M^{100} = \begin{pmatrix} 2^{100} & 3^{100} - 2^{100} \\ 0 & 3^{100} \end{pmatrix}.$$ □

Part Charlie: Review of complex numbers

For comparison, this part (not including the review) corresponds approximately to §11 of Poonen's notes.

§10 Complex numbers

I actually don't know why this subject is part of 18.02.

TODO: To be written

§11 Challenge review problems up to Midterm Exam 1

This is a set of six more difficult problems that I crafted for my students to help them prepare for their first midterm exam (which covered all the linear algebra parts and complex numbers). You can try them here if you want, but don't be discouraged if you find the problems tricky. All of these are much harder than anything that showed up on their actual midterm.

Suggested usage: think about each for 15-30 minutes, then read the solution. I hope this helps you digest the material; I tried to craft problems that teach deep understanding and piece together multiple ideas, rather than just using one or two isolated recipes. Solutions to these six problems are in Section 26.

Problem 11.1: In $\mathbb{R}^3$, compute the projection of the vector $\begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix}$ onto the plane $x + y + 2z = 0$.

Problem 11.2: Suppose A, B, C, D are points in $\mathbb{R}^3$. Give a geometric interpretation for this expression:

$$\left|\overrightarrow{DA} \cdot \left(\overrightarrow{DB} \times \overrightarrow{DC}\right)\right|.$$

Problem 11.3: Fix a plane $\mathcal{P}$ in $\mathbb{R}^3$ which passes through the origin. Consider the linear transformation $f : \mathbb{R}^3 \to \mathbb{R}^3$ where $f(\mathbf{v})$ is the projection of $\mathbf{v}$ onto $\mathcal{P}$. Let M denote the 3×3 matrix associated to f. Compute the determinant of M.

Problem 11.4: Let $\mathbf{a}$ and $\mathbf{b}$ be two perpendicular unit vectors in $\mathbb{R}^3$. A third vector $\mathbf{v}$ in $\mathbb{R}^3$ lies in the span of $\mathbf{a}$ and $\mathbf{b}$. Given that $\mathbf{v} \cdot \mathbf{a} = 2$ and $\mathbf{v} \cdot \mathbf{b} = 3$, compute the magnitudes of the cross products $\mathbf{v} \times \mathbf{a}$ and $\mathbf{v} \times \mathbf{b}$.

Problem 11.5: Compute the trace of the 2×2 matrix M given the two equations

$$M\begin{pmatrix} 4 \\ 7 \end{pmatrix} = \begin{pmatrix} 5 \\ 9 \end{pmatrix} \quad \text{and} \quad M\begin{pmatrix} 5 \\ 9 \end{pmatrix} = \begin{pmatrix} 4 \\ 7 \end{pmatrix}.$$

Problem 11.6: There are three complex numbers z satisfying $z^3 = 5 + 6i$. Suppose we plot these three numbers in the complex plane. Compute the area of the triangle they enclose.

Part Delta: Parametric side-quest

For comparison, this part corresponds approximately to §5 and §7 of Poonen's notes.

§12 Parametric equations

§12.1 [TEXT] Multivariate domains vs multivariate codomains

In 18.01, you did calculus on functions $F : \mathbb{R} \to \mathbb{R}$. So "multivariable calculus" could mean one of two things to start:

- Work with $F : \mathbb{R} \to \mathbb{R}^n$ instead (i.e. make the codomain multivariate).
- Work with $F : \mathbb{R}^n \to \mathbb{R}$ instead (i.e. make the domain multivariate).

What you should know now is **the first thing is WAY easier than the second**. This Part Delta is thus really short.

§12.2 [TEXT] Parametric pictures

From now on, we're going to usually change notation

$$\mathbf{r} : \mathbb{R} \to \mathbb{R}^n$$
$$\mathbf{r}(t) = \begin{pmatrix} \text{function in } t \\ \vdots \\ \text{function in } t \end{pmatrix}.$$

The choice of letter t for the input variable usually means "time"; and we use $\mathbf{r}$ for the function name to remind that the output is a vector.

</> Type signature

When you see $\mathbf{r}(t)$ or similar notation, the time variable t has type scalar, and the output is a vector.

If you're drawing a picture of a parametric function, usually all the axes are components of $\mathbf{r}(t)$ and the time variable doesn't have an axis. In other words, in the picture, **all the axis variables are output components, and we treat them all with equal respect**. The input time variable doesn't show up at all. (This is in contrast to 18.01 xy-graphs, where one axis was input and one axis was output. In the next section when we talk about *level curves*, it will be the other way around, where the output variable is anonymous and every axis is an input variable we treat with equal respect.)

Example

The classic example

$$\mathbf{r}(t) = \begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix}$$

would be drawn as the unit circle circle. You can imagine a particle starting at $\mathbf{r}(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and then moving around the unit circle counterclockwise with constant speed. It completes a full revolution in 2π time: $\mathbf{r}(2\pi) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$.

§12.3 [TEXT] Just always use components

Why is $\mathbf{r} : \mathbb{R} \to \mathbb{R}^n$ so easy that Part Delta is one section? Because there's pretty much only one thing you need to ever do:

> **Memorize**
>
> TLDR Just always use components.

That is, if $\mathbf{r} : \mathbb{R} \to \mathbb{R}^3$ (say), basically 90%+ of the time what you do is write

$$\mathbf{r}(t) = \langle x(t), y(t), z(t) \rangle = x(t)\mathbf{e}_1 + y(t)\mathbf{e}_2 + z(t)\mathbf{e}_3$$

and then just do single-variable calculus or calculations on each f_i.

- Need to differentiate $\mathbf{r}$? Differentiate each component.
- Need to integrate $\mathbf{r}$? Integrate each component.
- Need the absolute value of $\mathbf{r}$? Square root of sum of squares of components.

And so on. An example of Evan failing to do this is shown in Figure 8.

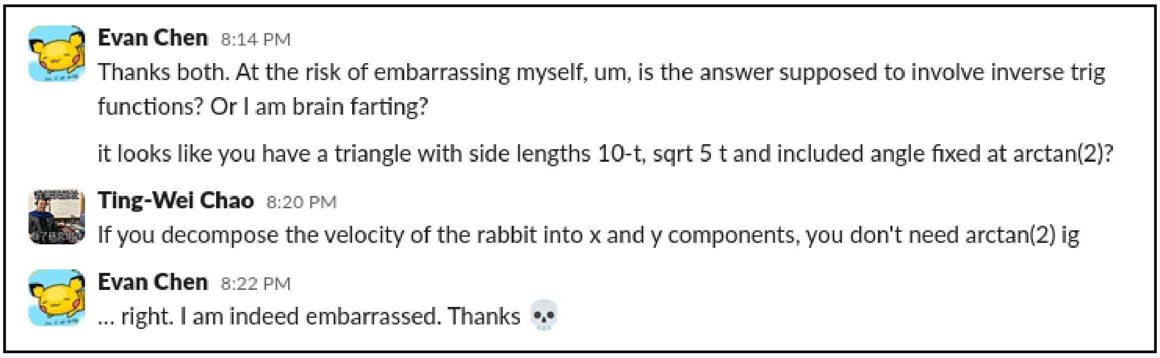

Figure 8: Seriously, just do everything componentwise.

§12.4 [RECIPE] Parametric things

I'll write this recipe with two variables, but it works equally well for three. Suppose you're given an equation $\mathbf{r}(t) = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}$. There are some things you could be asked:

> **Recipes for parametric stuff**
>
> - To find the **velocity vector** at a time t, it's the derivative
>
> $$\mathbf{r}'(t) = \begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix}.$$
>
> - To find the **speed** at a time t, it's the absolute value of the velocity:
>
> $$|\mathbf{r}'(t)| = \sqrt{x'(t)^2 + y'(t)^2}.$$
>
> - To find the **acceleration vector** at a time t, it's the second derivative of each component:
>
> $$\mathbf{r}''(t) = \begin{pmatrix} x''(t) \\ y''(t) \end{pmatrix}.$$

I don't know if there's a word for the absolute value of the acceleration vector (the way speed is the absolute value of the velocity vector).

Recipe for parametric integration

- To integrate $\mathbf{r}(t)$ between two times, take the integral of each component:

$$\int_{\text{start time}}^{\text{stop time}} \mathbf{r}(t)\,\mathrm{d}t = \begin{pmatrix} \int_{\text{start time}}^{\text{stop time}} x(t)\,\mathrm{d}t \\ \int_{\text{start time}}^{\text{stop time}} y(t)\,\mathrm{d}t \end{pmatrix}.$$

Recipe for arc length

The **arc length** from time t_{start} to t_{stop} is the integral of the speed:

$$\text{arc length} = \int_{\text{start time}}^{\text{stop time}} |\mathbf{r}'(t)|\,\mathrm{d}t.$$

(Technically, I should use "definition" boxes rather than "recipe" boxes here, since these are really the *definition* of the terms involved, and the recipes are "use the definition verbatim".)

Type signature

- Velocity $\mathbf{r}'(t)$, acceleration $\mathbf{r}''(t)$, and integrals $\int \mathbf{r}(t)\,\mathrm{d}t$ are vectors.
- But speed $|\mathbf{r}'(t)|$ and arc length are scalars.

Sample Question

Let

$$\mathbf{r}(t) = \begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix}.$$

Calculate:

- The velocity vector at time $t = \frac{\pi}{3}$.
- The speed at time $t = \frac{\pi}{3}$.
- The acceleration vector at time $t = \frac{\pi}{3}$.
- The integral $\int_0^{\frac{\pi}{3}} \mathbf{r}(t)\,\mathrm{d}t$.
- The arc length from $t = 0$ to $t = \frac{\pi}{3}$.

Solution: Let $\mathbf{r}(t) = \begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix}$. We will compute the following quantities.

Velocity vector at $t = \frac{\pi}{3}$ The velocity vector is the derivative of the position vector $\mathbf{r}(t)$ with respect to t:

$$\mathbf{v}(t) = \mathbf{r}'(t) = \begin{pmatrix} -\sin(t) \\ \cos(t) \end{pmatrix}.$$

At $t = \frac{\pi}{3}$, we have:

$$\mathbf{v}\left(\frac{\pi}{3}\right) = \begin{pmatrix} -\sin\left(\frac{\pi}{3}\right) \\ \cos\left(\frac{\pi}{3}\right) \end{pmatrix} = \begin{pmatrix} -\frac{\sqrt{3}}{2} \\ \frac{1}{2} \end{pmatrix}.$$

Thus, the velocity vector at $t = \frac{\pi}{3}$ is:

$$\mathbf{v}\left(\frac{\pi}{3}\right) = \begin{pmatrix} -\frac{\sqrt{3}}{2} \\ \frac{1}{2} \end{pmatrix}.$$

Speed at $t = \frac{\pi}{3}$ The speed is the magnitude of the velocity vector:

$$|\mathbf{v}(t)| = \sqrt{(-\sin(t))^2 + (\cos(t))^2} = \sqrt{\sin^2(t) + \cos^2(t)} = 1.$$

Thus, the speed at $t = \frac{\pi}{3}$ (or in fact any time) is:

$$\left|\mathbf{v}\left(\frac{\pi}{3}\right)\right| = 1.$$

Acceleration vector at $t = \frac{\pi}{3}$ Differentiate the velocity vector we got earlier:

$$\mathbf{a}(t) = \mathbf{v}'(t) = \begin{pmatrix} -\cos(t) \\ -\sin(t) \end{pmatrix}.$$

At $t = \frac{\pi}{3}$, we have:

$$\mathbf{a}\left(\frac{\pi}{3}\right) = \begin{pmatrix} -\cos\left(\frac{\pi}{3}\right) \\ -\sin\left(\frac{\pi}{3}\right) \end{pmatrix} = \begin{pmatrix} -\frac{1}{2} \\ -\frac{\sqrt{3}}{2} \end{pmatrix}.$$

Thus, the acceleration vector at $t = \frac{\pi}{3}$ is:

$$\mathbf{a}\left(\frac{\pi}{3}\right) = \begin{pmatrix} -\frac{1}{2} \\ -\frac{\sqrt{3}}{2} \end{pmatrix}.$$

Integral The integral of $\mathbf{r}(t)$ is computed component-wise:

$$\int_0^{\frac{\pi}{3}} \mathbf{r}(t)\,\mathrm{d}t = \int_0^{\frac{\pi}{3}} \begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix} \mathrm{d}t = \begin{pmatrix} \int_0^{\frac{\pi}{3}} \cos(t)\,\mathrm{d}t \\ \int_0^{\frac{\pi}{3}} \sin(t)\,\mathrm{d}t \end{pmatrix}.$$

Compute the integrals, using $\int \cos(t)\,\mathrm{d}t = \sin(t) + C$ and $\int \sin(t)\,\mathrm{d}t = \cos(t) + C$:

$$\int_0^{\frac{\pi}{3}} \cos(t)\,\mathrm{d}t = \sin\left(\frac{\pi}{3}\right) - \sin(0) = \frac{\sqrt{3}}{2}$$

$$\int_0^{\frac{\pi}{3}} \sin(t)\,\mathrm{d}t = -\cos\left(\frac{\pi}{3}\right) + \cos(0) = -\frac{1}{2} + 1 = \frac{1}{2}.$$

Thus, the integral is:

$$\int_0^{\frac{\pi}{3}} \mathbf{r}(t)\,\mathrm{d}t = \begin{pmatrix} \frac{\sqrt{3}}{2} \\ \frac{1}{2} \end{pmatrix}.$$

Arc length The arc length of a parametric curve is given by:

$$L = \int_0^{\frac{\pi}{3}} |\mathbf{r}'(t)|\,\mathrm{d}t = \int_0^{\frac{\pi}{3}} 1\,\mathrm{d}t = \frac{\pi}{3}.$$

Thus, the arc length from $t = 0$ to $t = \frac{\pi}{3}$ is:

$$L = \frac{\pi}{3}.$$

□

§12.5 [TEXT] Adding two vectors

Since everything is so mechanical once you have an equation for $\mathbf{r}(t)$, there's a shape of exam question that comes up in 18.02 where you're given some weird-looking path and need to get its equation in order to unlock things like velocity/speed/etc.

Something like 90%+ of the time if the shape is weird it's because it's the sum of two other vectors and you just add them.

The cycloid you saw in class was one hard-ish example of this. The curve looked scary. But you just ignore the shape, and just think about the equation

$$\mathbf{r}(t) = \langle tv, a\rangle + \langle a\cos\theta(t), a\sin\theta(t)\rangle.$$

Working out the angle is a bit annoying; but the point is no calculus or theory is involved, just work out the geometry. Then when you want the velocity, just differentiate $\mathbf{r}(t)$, and so on.

TODO: flesh this section out

§12.6 [TEXT] Eliminating the parameter t

For two-dimensional parametric pictures, the other shape of question that occasionally pops up is to get rid of t.

TODO: flesh this section out

Part Echo: Multivariable differentiation

For comparison, this part corresponds approximately to §8 and §12.1-§12.3 of Poonen's notes.

§13 Level curves (aka contour plots)

§13.1 [TEXT] Level curves replace xy-graphs

In high school and 18.01, you were usually taught to plot single-variable functions in two dimensions, so $f(x) = x^2$ would be drawn as a parabola $y = x^2$, and so on. You may have drilled into your head that x is an input and y is an output.

However, for 18.02 we'll typically want to draw pictures of functions like $f(x, y) = x^2 + y^2$ in a different way[9], using what's known as a *level curve*.

Definition

For any number c and function $f(x, y)$ the level curve for the value c is the plot of points for which $f(x, y) = c$.

The contrast to what you're used to is that:

- In high school and 18.01, the variables x and y play different roles, with x representing the input and $y = f(X)$ representing output.
- In 18.02, when we draw a function $f(x, y)$ both x and y are inputs; we treat them all with equal respect. Meanwhile, the *output* of the function does *not* have a variable name. If we really want to refer to it, we might sometimes write $f = 2$ as a shorthand for "the level curve for output 2".

To repeat that in table format:

18.01 xy-graphs	18.02 level curves
x is input	Both variables are inputs
y is output	No variable name for output

We give some examples.

[9]This is a lot like how we drew planes in a symmetric form earlier. In high school algebra, you drew 2D graphs of one-variable functions like $y = 2x + 5$ or $y = x^2 + 7$. So it might have seemed a bit weird to you that we wrote planes instead like $2x + 5y + 3z = 7$ rather than, say, $z = \frac{7-2x-5y}{3}$. But this form turned out to be better, because it let us easily access the normal vector (which here is $\langle 2, 5, 3 \rangle$). The analogy carries over here.

Example: the level curves of $f(x, y) = y - x^2$

To draw the level curves of the function $f(x, y) = y - x^2$, we begin by recalling that a level curve corresponds to a the points (x, y) such that the function takes on a constant value, say c. For our function, this becomes:

$$y - x^2 = c$$

which rearranges to

$$y = x^2 + c.$$

This is an equation in 18.01 form, where y is a function of x, so you can draw it easily. This equation represents a family of parabolas, each corresponding to a different value of c. As c varies, the level curves are parabolas that shift upward or downward along the y-axis. The shape of these curves is determined by the quadratic term x^2, which indicates that all level curves will have the same basic "U" shape.

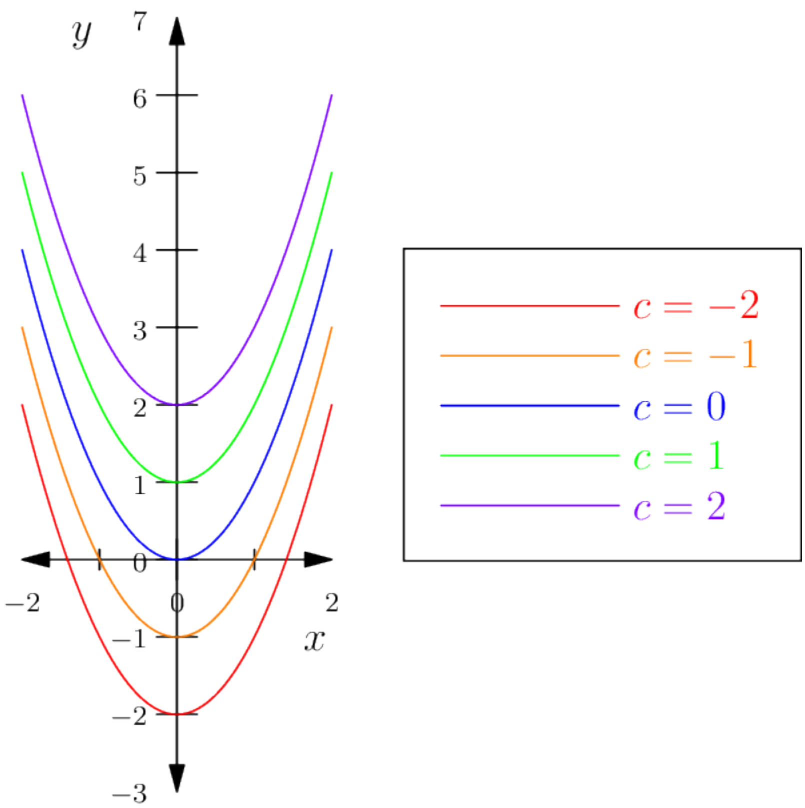

Figure 9: The level curves of $f(x, y) = y - x^2$.

Example: the level curves of $f(x,y) = x - y^2$

Let's draw level curves for $f(x,y) = x - y^2$. This example is exactly like the previous one, except the roles of x and y are flipped.

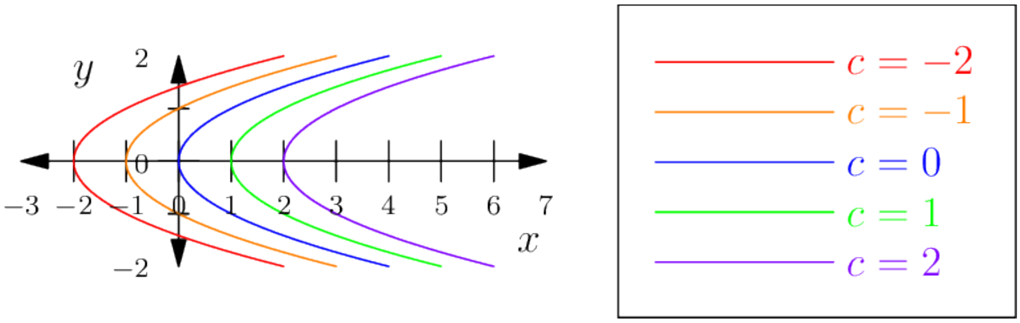

Figure 10: The level curves of $f(x,y) = x - y^2$.

Example: the level curves of $f(x,y) = x^2 + y^2$

Let's draw level curves of $f(x,y) = x^2 + y^2$ For each c we want to sketch the curve

$$x^2 + y^2 = c.$$

When $c < 0$, no points at all appear on this curve, and when $c = 0$ the only point is the origin $(0,0)$. For $c > 0$ this equation represents a family of circles centered at the origin $(0,0)$, with radius $\sqrt{c}$. For example:

- No points work for $c < 0$.
- For $c = 0$, the level curve is the single point $(0,0)$.
- For $c = 1$, the level curve is a circle with radius 1.
- For $c = 4$, the level curve is a circle with radius 2.
- For $c = 9$, the level curve is a circle with radius 3.

As c increases, the circles expand outward from the origin. These concentric circles represent the level curves of the function $f(x,y) = x^2 + y^2$.

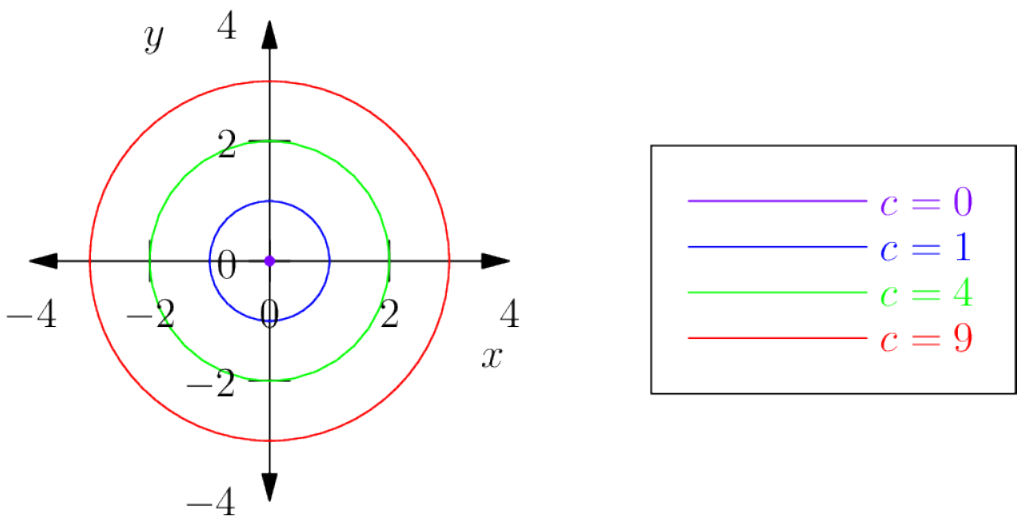

Figure 11: Four of the level curves for $f(x,y) = x^2 + y^2$.

Example: the level curves of $f(x,y) = |x| + |y|$

Let's draw level curves of $f(x,y) = |x| + |y|$. To draw the level curve for c, we are looking at

$$|x| + |y| = c.$$

Like before, if $c < 0$ there are no pairs (x,y) at all and for $c = 0$ there is only a single point.

This equation represents a family of polygons. Specifically, for a given value of c, the points that satisfy this equation form a diamond shape centered at the origin. Indeed, in the first quadrant (where the absolute values don't do anything) it represents the line segment joining $(0,c)$ to $(c,0)$.

So for example,

- When $c < 0$, there are no points.
- For $c = 0$, the level curve is just the point $(0,0)$.
- For $c = 1$, the level curve is a diamond with vertices at $(1,0)$, $(-1,0)$, $(0,1)$, and $(0,-1)$.
- For $c = 2$, the level curve is a larger diamond with vertices at $(2,0)$, $(-2,0)$, $(0,2)$, and $(0,-2)$.
- For $c = 3$, the diamond expands further, with vertices at $(3,0)$, $(-3,0)$, $(0,3)$, and $(0,-3)$.

As c increases, the diamonds expand outward, maintaining their shape but increasing in size. Each level curve is a square rotated by 45 degrees.

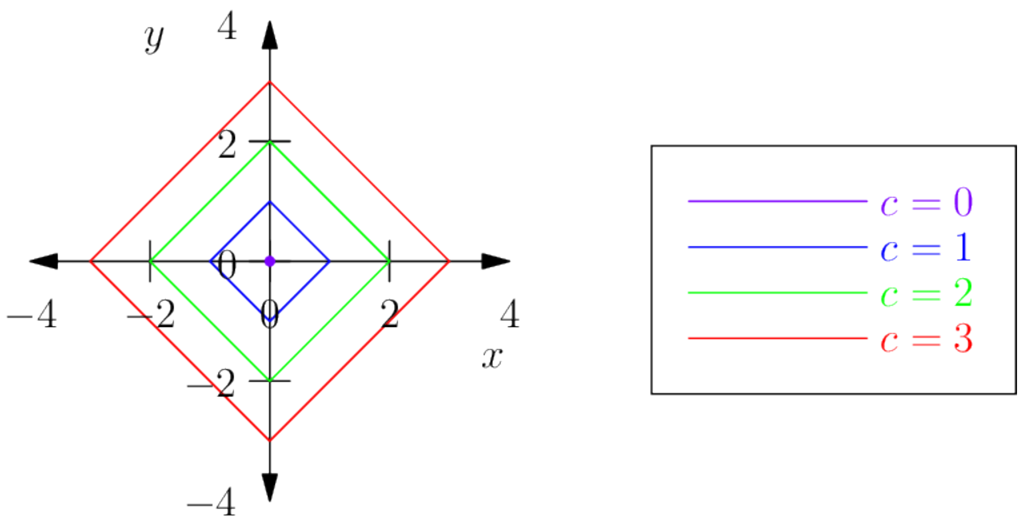

Figure 12: Four of the level curves for $f(x,y) = |x| + |y|$.

§13.2 [RECIPE] Drawing level curves

Despite the fact this section is labeled "recipe", there isn't an easy method that works for every function. **You have to do it in an ad-hoc way depending on the exact function you're given**. For many functions you'll see on an exam, it'll be pretty easy.

To summarize the procedure, given an explicit function like $f(x,y)$ and the value of c, one tries to plot all the points (x,y) in space with $f(x,y) = c$. We gave three examples right above, where:

- The level curves of $f(x,y) = y - x^2$ were easy to plot because for any given c, the equation just became an xy-plot like in 18.01.

- The level curves of $f(x,y) = x - y^2$ were similar to the previous example, but the roles of x and y were flipped.
- To draw the level curves of $f(x,y) = x^2 + y^2$, you needed to know that $x^2 + y^2 = r^2$ represents a circle of radius r centered at $(0,0)$.
- To draw the level curves of $f(x,y) = |x| + |y|$, we had to think about it an ad-hoc manner where we worked in each quadrant; in Quadrant I we figured out that we got a line, and then we applied the same image to the other quadrants to get diamond shapes.

So you can see it really depends on the exact f you are given. If you wrote a really nasty function like $f(x,y) = e^{\sin xy} + \cos(x+y)$, there's probably no easy way to draw the level curve by hand.

§13.3 [TEXT] Level surfaces are exactly the same thing, with three variables instead of two

Nothing above really depends on having exactly two variables. If we had a three-variable function $f(x,y,z)$, we could draw *level surfaces* for a value of c by plotting all the points in $\mathbb{R}^3$ for which $f(x,y,z) = c$.

Example: Level surface of $f(x,y,z) = x^2 + y^2 + z^2$

If $f(x,y,z) = x^2 + y^2 + z^2$, then the level surface for the value c will be a sphere with radius $\sqrt{c}$ if $c \geq 0$. (When $c < 0$, the level surface is empty.)

Example: Level surface of $f(x,y,z) = x + 2y + 3z$

If $f(x,y,z) = x + 2y + 3z$, all the level surfaces of f are planes in $\mathbb{R}^3$, which are parallel to each other with normal vector $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$.

§13.4 [EXER] Exercises

Exercise 13.1: Draw 2-D level curves for some values for the following functions:

- $f(x,y) = 5x + y$
- $f(x,y) = xy$
- $f(x,y) = e^{x^2+y^2}$
- $f(x,y) = \max(x,y)$ (i.e. f outputs the larger of its two inputs, so $f(3,5) = 5$ and $f(2,-9) = 2$, for example).

*** Exercise 13.2**: Give an example of a polynomial function $f(x,y)$ for which the level curve for the value 100 consists of exactly seven points.

§14 Partial derivatives

§14.1 [TEXT] The point of differentiation is linear approximation

In 18.01, when $f : \mathbb{R} \to \mathbb{R}$, you defined a **derivative** $f'(p)$ at each input $p \in \mathbb{R}$, which you thought of as the **slope** of the **tangent line** at p. Think $f(5.01) \approx f(5) + f'(5) \cdot 0.01$. This slope roughly tells you, if you move a slight distance away from the input p, this is how fast you expect f to change. To drill the point home again, in 18.01, we had

$$f(p + \varepsilon) = f(p) + f'(p) \cdot \varepsilon.$$

See figure below.

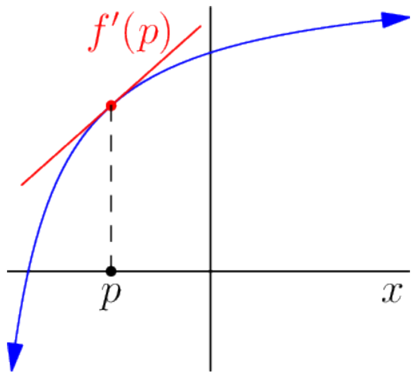

Figure 13: In 18.01, the slope $f'(p)$ tells you how quickly f changes near p.

The 18.01 derivative had type "scalar". But for a *two-variable* function, that's not enough. For concreteness, let's take

$$f(x, y) = x^2 + y^2$$

as our example function (for which we have drawn level curves before), and consider some point $P = (3, 4)$, so that $f(3, 4) = 25$.

Then, what would a point "close" to $(3, 4)$ mean? The point $(3.01, 4)$ is close, but so is $(3, 4.01)$. For that matter, so is $(3.006, 4.008)$ — that's also a point at distance 0.01 away! So having a single number isn't enough to describe the rate of change anymore.

For a two-variable function, we would really want *two* numbers, in the sense that we want to fill in the blanks in the equation

$$f\left(3 + \varepsilon_x, 4 + \varepsilon_y\right) \approx 25 + (\text{slope in } x\text{-direction}) \cdot \varepsilon_x + (\text{slope in } y\text{-direction}) \cdot \varepsilon_y.$$

> **Idea**
>
> For an n-variable functions, we have a rate of change in *each* of the n directions. Therefore, **we need n numbers and not just one**.

The first blank corresponds to what happens if you imagine y is held in place at 4, and we're just changing the x-value to 3.01. The second blank is similar. So we need a way to calculate these; the answer to our wish is what's called a *partial derivative*.

§14.2 [TEXT] Computing partial derivatives is actually just 18.01

The good news about partial derivatives is that **they're actually really easy to calculate**. You pretty much just need to do what you were taught in 18.01 with one variable changing while pretending the others are constants.

Here's the definition:

Definition

Suppose $f(x, y)$ is a two-variable function. Then the *partial derivative with respect to* x, which we denote either f_x or $\frac{\partial f}{\partial x}$, is the result if we differentiate f while treating x as a variable y as a constant. The partial derivative $f_y = \frac{\partial f}{\partial y}$ is defined the same way.

Similarly, if $f(x, y, z)$ is a three-variable function, we write $f_x = \frac{\partial f}{\partial x}$ for the derivative when y and z are fixed.

Type signature

Each partial derivative has the same type signature as f. That is:

- Given $f : \mathbb{R}^n \to \mathbb{R}$ which accepts **points** in $\mathbb{R}^n$ and outputs **scalars**.
- Then the partial derivative $\frac{\partial f}{\partial x} = f_x$ also accepts **points** in $\mathbb{R}^n$ and outputs **scalars**.

But that's a lot of words. I think this is actually better explained by example. In fact you could probably just read the examples and ignore the definition above.

Example: partial derivatives of $f(x, y) = x^3y^2 + \cos(y)$

Let $f(x, y) = x^3y^2 + \cos(y)$.

Let's compute f_x. Again, pretend y is a constant, so look at the function

$$x \mapsto y^2 \cdot x^3 + \cos(y).$$

If we differentiate with respect to x, then x^3 becomes $3x^2$, and $\cos(y)$ goes to 0 (it doesn't have any x stuff in it). So

$$f_x = y^2 \cdot 3x^2.$$

Similarly, let's compute f_y. This time we pretend x is a constant, and look at

$$y \mapsto x^3 \cdot y^2 + \cos(y).$$

This time y^2 becomes $2y$, and $\cos(y)$ has derivative $-\sin(y)$. So

$$f_y = x^3 \cdot 2y - \sin(y).$$

Example: partial derivatives of $f(x,y,z) = e^{xyz}$

Let $f(x,y,z) = e^{xyz}$ for a three-variable example. To compute f_x, think of the function

$$x \mapsto e^{yz \cdot x}$$

where we pretend y and z are constants. Then the derivative is with respect to x is just $yze^{yz \cdot x}$ (just like how the derivative of e^{3x} is $3e^x$). In other words,

$$f_x(x,y,z) = yz \cdot e^{xyz}.$$

For analogous reasons:

$$f_y(x,y,z) = xz \cdot e^{xyz}$$
$$f_z(x,y,z) = xy \cdot e^{xyz}.$$

Example: partial derivatives of $f(x,y) = x^2 + y^2$ and linear approximation

Let's go back to

$$f(x,y) = x^2 + y^2$$

which we used in our earlier example as motivation, at the point $P = (3,4)$.

Let's fill in the numbers for the example $f(x,y) = x^2 + y^2$ we chose. By now, you should be able to compute that

$$f_x(x,y) = 2x$$
$$f_y(x,y) = 2y$$

Now, let's zoom in on just the point $P = (3,4)$. We know that

$$f(P) = 3^2 + 4^2 = 25$$
$$f_x(P) = 2 \cdot 3 = 6$$
$$f_y(P) = 2 \cdot 4 = 8.$$

So our approximation equation can be written as

$$(3 + \varepsilon_x)^2 + \left(4 + \varepsilon_y\right)^2 \approx 25 + 6\varepsilon_x + 8\varepsilon_y. \tag{3}$$

If you manually expand both sides, you can see this looks true. The two sides differ only by ε_x^2 and ε_y^2, and the intuition is that if ε_x and ε_y were small numbers, then their squares will be negligibly small.

We'll return to Equation 3 later when we introduce the gradient.

§14.3 [RECIPE] Computing partial derivatives

You probably can already figure out the recipe from the sections above, but let's write it here just for completeness.

Recipe for calculating partial derivatives

To compute the partial derivative of a function $f(x, y)$ or $f(x, y, z)$ or $f(x_1, ..., x_n)$ with respect to one of its input variables,

1. Pretend all the other variables are constants, and focus on just the variable you're taking the partial derivative to.
2. Calculate the derivative of f with respect to just that variable like in 18.01.
3. Output the derivative you got.

This is easy, and only requires 18.01 material.

We just saw three examples where we computed the partials for $f(x, y) = x^3y^2 + \cos(y)$, $f(x, y, z) = e^{xyz}$, and $f(x, y) = x^2 + y^2$. Here are a bunch more examples that you can try to follow along:

Sample Question

Calculate the partial derivatives of $f(x, y, z) = x + y + z$.

Solution: The partial derivative with respect to x is obtained by differentiating

$$x \mapsto x + y + z.$$

Since we pretend y and z are constants, we just differentiate x to get 1. The same thing happens with y and z. Hence

$$\begin{aligned} f_x(x, y, z) &= 1 \\ f_y(x, y, z) &= 1 \\ f_z(x, y, z) &= 1. \end{aligned}$$

□

Sample Question

Calculate the partial derivatives of $f(x, y, z) = xy + yz + zx$.

Solution: We differentiate with respect to x first, where we view as the function

$$x \mapsto (y + z)x + yz$$

pretending that y and z are constants. This gives derivative $f_x(x, y, z) = y + z$. Similarly, $f_y(x, y, z) = x + z$ and $f_z(x, y, z) = x + y$. So

$$\begin{aligned} f_x(x, y, z) &= y + z \\ f_y(x, y, z) &= z + x \\ f_z(x, y, z) &= x + y. \end{aligned}$$

□

Sample Question

Calculate the partial derivatives of $f(x, y) = x^y$, where we assume $x, y > 0$.

Solution: Our last example is

$$f(x, y) = x^y,$$

where let's say for $x, y > 0$ for simplicity (otherwise the exponentiation may not be defined).

If we view y as a constant and x as a variable, then

$$x \mapsto x^y$$

is differentiated by the "power rule" to get yx^{y-1}. However, if we view x as constant and y as a variable, then

$$y \mapsto x^y = e^{\log x \cdot y}$$

ends up with derivative $\log x \cdot e^{\log x \cdot y} = \log x \cdot e^y$. Hence

$$f_x(x, y) = yx^{y-1}$$
$$f_y(x, y) = \log x \cdot e^y.$$

□

§14.4 [EXER] Exercises

Exercise 14.1: Find all the partial derivatives of the following functions, defined for $x, y, z > 0$:

- $f(x, y, z) = \frac{x}{y} + \frac{y}{z} + \frac{z}{x}$
- $f(x, y, z) = \sin(xyz)$
- $f(x, y, z) = x^y + y^z + z^x$.

§15 The gradient

The gradient of $f : \mathbb{R}^n \to \mathbb{R}$, denoted ∇f, is the single most important concept in the entire "Multivariable differentiation" part. Although its definition is actually quite easy to compute, I want to give a proper explanation for where it comes from.

Throughout this section, remember two important ideas:

- The goal of the derivative is to approximate a function by a linear one.
- Everything you used slopes for before, you should use normal vectors instead.

If you want spoilers for what's to come, see the following table.

Thing	**18.01**	**18.02**
Input	$f : \mathbb{R} \to \mathbb{R}$	$f : \mathbb{R}^n \to \mathbb{R}$
Output	$f' : \mathbb{R} \to \mathbb{R}$	$\nabla f : \mathbb{R}^n \to \mathbb{R}^n$
Think of as	Slope (rise/run)	Measures change in *each* of n directions
Approximation	Multiply by small run	Dot product with small displacement
Picture	Slope of tangent in xy-graph	Normal vector to tangent of level curve

Table 3: How to think of ∇f for multivariable functions, compared to the derivative in 18.01.

§15.1 [TEXT] The gradient rewrites linear approximation into a dot product

In 18.01, when $f : \mathbb{R} \to \mathbb{R}$ was a function and $p \in \mathbb{R}$ was an input, we thought of the single number $f'(p)$ as the slope to interpret it geometrically. Now that we're in 18.02, we have n different rates of change, but we haven't talked about how to think of it geometrically yet.

It turns out correct definition is to take the n numbers and make them into a vector. Bear with me for just one second:

Definition

If $f(x, y)$ is a two-variable function (so $f : \mathbb{R}^2 \to \mathbb{R}$), the **gradient** of f, denoted ∇f, is the function $\mathbb{R}^2 \to \mathbb{R}^2$ obtained by taking the two partial derivatives as the coordinates:

$$\nabla f(x, y) = \begin{pmatrix} f_x(x, y) \\ f_y(x, y) \end{pmatrix}.$$

The case of n variables is analogous; for example if $f(x, y, z)$ is a three-variable function, then

$$\nabla f(x, y, z) = \begin{pmatrix} f_x(x, y, z) \\ f_y(x, y, z) \\ f_z(x, y, z) \end{pmatrix}.$$

</> Type signature

The types are confusing here. To continue harping on type safety:

- Suppose $f : \mathbb{R}^2 \to \mathbb{R}$ accepts **points** in $\mathbb{R}^2$ and outputs **scalars** in $\mathbb{R}$.
- Then $\nabla f : \mathbb{R}^2 \to \mathbb{R}^2$ accepts **points** in $\mathbb{R}^2$ and outputs **vectors** in $\mathbb{R}$.

Keep the distinction between points and vectors in mind when drawing pictures. We'll always draw points as dots, as arrows.

The reason for defining this gradient is that it lets us do linear approximation with a **dot product**, and consequently dot products are going to be super important throughout this section. Let me show you how. Let's go back to our protagonist

$$f(x, y) = x^2 + y^2$$

at the point $P = (3, 4)$. Way back in Equation 3 (on page 56), we computed $f_x(P) = 2 \cdot 3 = 6$ and $f_y(P) = 2 \cdot 4 = 8$ and used it to get the approximation

$$\begin{aligned} f(P + \langle \varepsilon_x, \varepsilon_y \rangle) &= f(\langle 3, 4 \rangle + \langle \varepsilon_x, \varepsilon_y \rangle) \\ &= (3 + \varepsilon_x)^2 + (4 + \varepsilon_y)^2 \approx 25 + 6\varepsilon_x + 8\varepsilon_y. \end{aligned}$$

Now the idea that will let us do geometry is to replace the pair of numbers ε_x and ε_y with a single "small displacement" vector $\mathbf{v} = \begin{pmatrix} \varepsilon_x \\ \varepsilon_y \end{pmatrix}$, and the pair of numbers 6 and 6 with the vector $\begin{pmatrix} 6 \\ 8 \end{pmatrix}$ instead, so that **the approximation part just becomes a dot product**:

$$f\left(\begin{pmatrix} 3 \\ 4 \end{pmatrix} + \mathbf{v}\right) \approx f\left(\begin{pmatrix} 3 \\ 4 \end{pmatrix}\right) + \begin{pmatrix} 6 \\ 8 \end{pmatrix} \cdot \mathbf{v}.$$

The approximation part is used so often it has its own name and symbol.

⚠ Warning

In some places you see the abbreviation $D_{\mathbf{v}} f(P) := \nabla f(P) \cdot \mathbf{v}$ and the name "directional derivative" for it. I hate this term, because some people have different notations and definitions (according to Wikipedia, some authors require $\mathbf{v}$ to be a unit vector, etc.).

So I will always just write the dot product $\nabla f(P) \cdot \mathbf{v}$ instead, which is unambiguous and means you have one less symbol to remember.

In full abstraction, we can rewrite linear approximation as:

! Memorize

Suppose f is differentiable at a point P. Then for small displacement vectors $\mathbf{v}$, the **net change** from $f(P)$ to $f(P + \mathbf{v})$ is approximated by the dot product

$$\nabla f(P) \cdot \mathbf{v}.$$

This procedure is called **linear approximation**.

Up until now, all we've done is rewrite the earlier equation with a different notation; so far, nothing new has been introduced. Why did we do all this work to use different symbols to say the same thing?

The important idea is what I told you a long time ago: **anything you used to think of in terms of slopes, you should rethink in terms of normal vectors**. It turns out that to complete the analogy to differentiation, the normal vector is going to be that gradient $\nabla f(P)$, and we'll see why in just a moment (spoiler: it's because of the dot product). For now, you should just know that $\nabla f(P)$ is *going to be* the right way to draw pictures of all n rates of change at once, although I haven't explained why yet.

Before going on, let's write down the recipes and some examples just to make sure the *definition* of the gradient makes sense, then I'll explain why the gradient is the normal vector we need to complete our analogy.

§15.2 [RECIPE] Calculating the gradient

Recipe for calculating the gradient

1. Compute every partial derivative of the given function.
2. Output the vector whose components are those partial derivatives.

Sample Question

Consider the six functions

$$\begin{aligned} f_1(x,y) &= x^3y^2 + \cos(y), & f_2(x,y,z) &= e^{xyz} \\ f_3(x,y) &= x^2 + y^2, & f_4(x,y,z) &= x + y + z \\ f_5(x,y,z) &= xy + yz + zx & f_6(x,y) &= x^y \end{aligned}$$

from back in Section 14.2 and Section 14.3. Compute their gradients.

Solution: Take the partial derivatives we already computed and make them the components:

$$\begin{aligned} \nabla f_1(x,y) &= \begin{pmatrix} 3x^2y^2 \\ 2x^3y - \sin(y) \end{pmatrix}, & \nabla f_2(x,y) &= \begin{pmatrix} yze^{xyz} \\ xze^{xyz} \\ xye^{xyz} \end{pmatrix}, \\ \nabla f_3(x,y) &= \begin{pmatrix} 2x \\ 2y \end{pmatrix}, & \nabla f_4(x,y,z) &= \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \\ \nabla f_5(x,y,z) &= \begin{pmatrix} y+z \\ x+z \\ x+y \end{pmatrix}, & \nabla f_6(x,y) &= \begin{pmatrix} yx^{y-1} \\ \log(y) \cdot x^y \end{pmatrix}. \end{aligned}$$

□

§15.3 [RECIPE] Linear approximation

We actually could have stated an equivalent recipe right after we defined partial derivatives, but conceptually I think it's better to think of everything in terms of the gradient, so I waited until after I had defined the gradient to write the recipe.

Recipe for linear approximation

To do linear approximation of $f(P + \mathbf{v})$ for a small displacement vector $\mathbf{v}$:

1. Compute $\nabla f(P)$, the gradient of f at the point P.
2. Take the dot product $\nabla f(P) \cdot \mathbf{v}$ to get a number, the approximate change.
3. Output $f(P)$ plus the change from the previous step.

Sample Question

Let $f(x, y) = x^2 + y^2$. Approximate the value of $f(3.01, 4.01)$ by using linear approximation from $(3, 4)$.

Solution: Compute the gradient by taking both partial derivatives:

$$\nabla f(x, y) = \begin{pmatrix} 2x \\ 2y \end{pmatrix}.$$

So the gradient vector at the starting point is given by

$$\nabla f(3, 4) = \begin{pmatrix} 2 \cdot 3 \\ 2 \cdot 4 \end{pmatrix} = \begin{pmatrix} 6 \\ 8 \end{pmatrix}.$$

The target point $(3.01, 4.01)$ differs from the starting point $(3, 4)$ by the displacement $\mathbf{v} = (0.01, 0.01)$. So the approximate change in f is given by

$$\underbrace{\begin{pmatrix} 6 \\ 8 \end{pmatrix}}_{= \nabla f(3,4)} \cdot \underbrace{\begin{pmatrix} 0.01 \\ 0.01 \end{pmatrix}}_{= \mathbf{v}} = (6 \cdot 0.01 + 8 \cdot 0.01) = 0.14.$$

Therefore,

$$f(3.01, 4.01) \approx \underbrace{f(3, 4)}_{=25} + 0.14 = 25.14.$$

□

Sample Question

Let $f(x, y) = x^3 - y^3$. Approximate the value of $f(2.01, -1.01)$ by using linear approximation from $(2, -1)$.

Solution: Compute the gradient by taking both partial derivatives:

$$\nabla f(x, y) = \begin{pmatrix} 3x^2 \\ -3y^2 \end{pmatrix}.$$

So the gradient vector at the starting point $(2, -1)$ is given by

$$\nabla f(2, -1) = \begin{pmatrix} 3(2)^2 \\ -3(-1)^2 \end{pmatrix} = \begin{pmatrix} 12 \\ -3 \end{pmatrix}.$$

The target point $(2.01, -1.01)$ differs from the starting point $(2, -1)$ by the displacement $\mathbf{v} = (0.01, -0.01)$. So the approximate change in f is given by

$$\underbrace{\begin{pmatrix} 12 \\ -3 \end{pmatrix}}_{=\nabla f(2,-1)} \cdot \underbrace{\begin{pmatrix} 0.01 \\ -0.01 \end{pmatrix}}_{=\mathbf{v}} = (12 \cdot 0.01 + (-3) \cdot (-0.01)) = 0.15.$$

Therefore,

$$f(2.01, -1.01) \approx \underbrace{f(2,-1)}_{=9} + 0.15 = 9.14.$$

□

Sample Question

Let $f(x, y) = e^x \sin(y) + 777$. Approximate the value of $f(0.04, 0.03)$ by using linear approximation from the point $(0, 0)$.

Solution: Compute the gradient by taking both partial derivatives:

$$\nabla f(x, y) = \begin{pmatrix} e^x \sin y \\ e^x \cos y \end{pmatrix}.$$

So the gradient vector at the starting point $(0, 0)$ is given by

$$\nabla f(0, 0) = \begin{pmatrix} e^0 \sin 0 \\ e^0 \cos 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

The target point $(0.04, 0.03)$ differs from the starting point $(0, 0)$ by $(0.04, 0.03)$. So the approximate change in f is given by

$$\underbrace{\begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{=\nabla f(0,0)} \cdot \underbrace{\begin{pmatrix} 0.04 \\ 0.03 \end{pmatrix}}_{=\mathbf{v}} = 0 \cdot 0.04 + 1 \cdot 0.03 = 0.04.$$

Therefore,

$$f(0.04, 0.03) \approx \underbrace{f(0,0)}_{=777} + 0.03 = 777.03.$$

□

§15.4 [TEXT] Gradient descent

At the end of Section 15.1, we promised the geometric definition of the dot product would pay dividends. We now make good on that promise.

The motivating question here is:

Question

Let $f(x, y) = x^2 + y^2$. Imagine we're standing at the point $P = (3, 4)$. We'd like to take a step 0.01 away in some direction of our choice. For example, we could go to $(2.99, 4)$, or $(3, 4.01)$ or $(2.992, 4.006)$, or any other point on the circle we've marked in the figure below. (For the third point, note that $\sqrt{(3 - 2.992)^2 - (4.006)^2} = 0.01$, so that point is indeed 0.01 away.)

- Which way should we step if we want to maximize the f-value at the new point?
- Which way should we step if we want to the f-value to stay about the same?
- Which way should we step if we want to minimize the f-value at the new point?

You can see a cartoon of the situation in Figure 14. Note that this figure is not to scale, because 0.01 is too small to be legibly drawn, so the black circle is drawn much larger than it actually is.

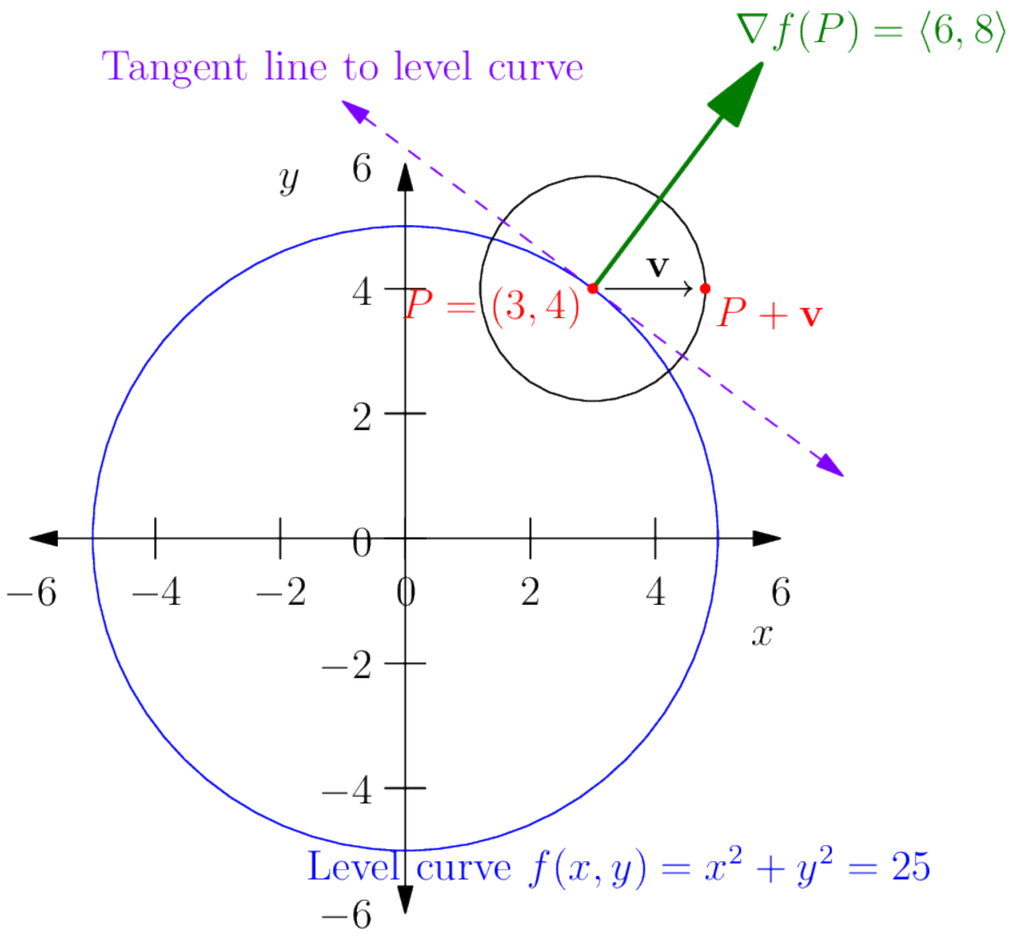

Figure 14: Starting from $P = (3, 4)$, we make a step $\mathbf{v}$ away, where $|\mathbf{v}| = 0.01$. Not to scale.

To answer the question, we use the geometric interpretation of the dot product now. Remember that the change in f is approximated by

$$f(P + \mathbf{v}) - f(P) \approx \nabla f(P) \cdot \mathbf{v}.$$

The geometric definition of the dot product is that it equals

$$\nabla f(P) \cdot \mathbf{v} = |\nabla f(P)| \, |\mathbf{v}| \cos \theta$$

where θ is the included angle. But $|\nabla f(P)|$ is fixed (in this example, it's $\sqrt{6^2 + 8^2} = 10$) and $|\mathbf{v}|$ is fixed as well (in this example we chose it to be the small number 0.01).

So actually all we care about is the angle θ! Think about that for a moment. Then remember how the cosine function works:

- $\cos(0°) = 1$ is the most positive value of the cosine, and that occurs when $\mathbf{v}$ and $\nabla f(P)$ point the same direction.
- $\cos(180°) = 1$ is the most negative value of the cosine, and that occurs when $\mathbf{v}$ and $\nabla f(P)$ point the same direction.
- If $\nabla f(P)$ and $\mathbf{v}$ are perpendicular (so $\theta = 90°$ or $\theta = 270°$), then the dot product is zero.

Translation:

Memorize

- Move **along** the gradient to increase f as quickly as possible.
- Move **against** the gradient to decrease f as quickly as possible.
- Move **perpendicular to** the gradient to avoid changing f by much either direction.

§15.5 [TEXT] Normal vectors to the tangent line/plane

We only need to add one more idea: *keeping f about the same should correspond to moving along the tangent line or plane.*

Indeed, in the 2D case, the tangent line is the line that "hugs" the level curve the closest, so we think of it as the direction causing f to avoid much change. The same is true for a tangent plane to a level surface in the 3D case; the plane hugs the curve near the point P. So that means the last bullet could be rewritten as

Memorize

The gradient $\nabla f(P)$ is normal to the tangent line/plane at P. It points towards the direction that increases f.

Example

In the previous example with a level curve, the gradient pointed away from the interior. This is not true in general. For example, imagine instead the function

$$f(x, y) = \frac{1}{x^2 + y^2}.$$

The point $(3, 4)$ lies on the level curve of $f(3, 4) = \frac{1}{25}$. The level curve of $f(x, y)$ with value $\frac{1}{25}$ is *also* a circle of radius 5, because it corresponds to the equation $\frac{1}{x^2+y^2} = \frac{1}{25}$.

However, the gradient looks quite different: with enough calculation one gets

$$\nabla f(x, y) = \begin{pmatrix} \frac{-2x}{(x^2+y^2)^2} \\ \frac{-2y}{(x^2+y^2)^2} \end{pmatrix}.$$

Evaluating at $(3, 4)$, we get

$$\nabla f(3, 4) = \begin{pmatrix} -\frac{6}{625} \\ -\frac{8}{625} \end{pmatrix}.$$

Hence, for the function $f(x, y) = \frac{1}{x^2+y^2}$, drawing the figure analogous to Figure 14 gives something that looks quite similar, except the green arrow points the *other* way and is way smaller. This makes sense: as you move *towards* the origin, you expect $\frac{1}{x^2+y^2}$ to get larger. See Figure 15.

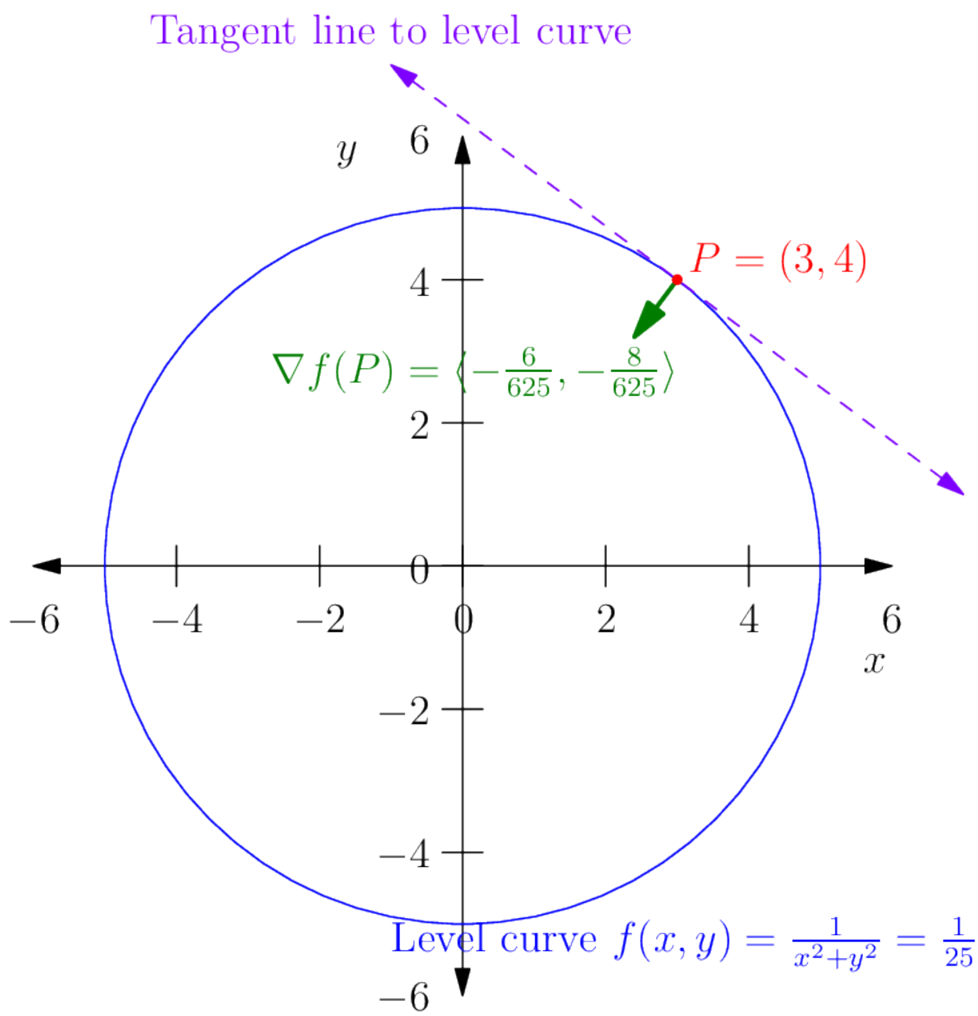

Figure 15: Similar picture but for $f(x, y) = \frac{1}{x^2+y^2}$. It looks very similar to Figure 14, but now the gradient points the other way and has much smaller absolute value, indicating that the value of f increases as we go *towards* the center (but only slightly). Not to scale.

> **Remark**
>
> Back in the 3D geometry in the linear algebra part of the course, we usually neither knew nor cared what the sign and magnitude of the normal vector was. That is, when asked "what is a normal vector to the plane $x - y + 2z = 8$?", you could answer $\begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}$ or $\begin{pmatrix} -1 \\ 1 \\ -2 \end{pmatrix}$ or even $\begin{pmatrix} -100 \\ 100 \\ -200 \end{pmatrix}$. But this doesn't apply to the gradient anymore: while it is a normal vector to the tangent line/plane, the magnitude carries additional information we shouldn't just throw away.

§15.6 [RECIPE] Computing tangent lines/planes to level curves/surfaces

At this point, we can compute tangent lines and planes easily. We apply the old recipe in Section 5.4 (finding a plane given a point with a known normal vector) with $\nabla f(P)$ as the normal vector. To spell it out:

Recipe: Tangent line/plane to level curve/surface

To find the tangent line/plane to a level curve/surface of a function f at point P:

1. Compute the gradient ∇f. This is a normal vector, so it tells you the left-hand side for the equation of the line/plane.
2. Adjust the right-hand side so it passes through P, like in Section 5.4.

Sample Question

Find the tangent line to $x^2 + y^2 = 25$ at the point $(3, 4)$.

Solution: Let $f(x, y) = x^2 + y^2$, so we are looking at the level curve for 25 of f. We have seen already that

$$\nabla f = \begin{pmatrix} 6 \\ 8 \end{pmatrix}.$$

Hence, the tangent line should take the form

$$6x + 8y = d$$

for some d. To pass through $P = (3, 4)$, we need $d = 42$, so the answer is

$$6x + 8y = 42.$$

□

TODO: A couple more examples here would be nice...

§15.7 [RECAP] A recap of Part Echo on Multivariable Differentation

Let's summarize the last few sections.

- We replaced the old graphs we used in 18.01 with level curve and level surface pictures in Section 13. These new pictures differed from 18.01 pictures because all the variables on the axes are inputs now, and we treat them all with equal respect.
- We explained in Section 14 how to take a partial derivative of $f(x, y)$ or $f(x, y, z)$, which measures the change in just one of the variables.
- We used these partial derivatives to define the gradient ∇f in Section 15. This made linear approximation into a dot product, where $f(P + \mathbf{v}) \approx f(P) + \nabla f(P) \cdot \mathbf{v}$ for a small displacement $\mathbf{v}$.
- Using the geometric interpretation of a dot product, $\nabla f(P)$ was a normal vector to the level curve of f passing through P, and:
  - Going along the gradient increases f most rapidly
  - Going against the gradient decreases f most rapidly
  - Going perpendicular to the gradient puts you along the tangent line or plane at P.

§15.8 [EXER] Exercises

Exercise 15.1: Find the tangent plane to the sphere $x^2 + y^2 + z^2 = 14$ at the point $(1, 2, 3)$.

Exercise 15.2: The level curve of a certain function $f(x, y)$ for the value -7 turns out to be a circle of radius 2 centered at $(0, 0)$.

- What are all possible vectors that $\nabla f(1.2, -1.6)$ could be?
- Do linear approximation to estimate $f(1.208, -1.594)$ starting from the point $(1.2, -1.6)$.

Part Foxtrot: Optimization

For comparison, this part corresponds approximately to §9 and §12.4-§12.6 of Poonen's notes.

§16 Critical points

§16.1 [TEXT] Critical points in 18.01

First, a comparison to 18.01. Way back when you had a differentiable single-variable function $f : \mathbb{R} \to \mathbb{R}$, and you were trying to minimize it, you used the following terms:

18.01 term	**Meaning**
Global minimum	Minimum of the function f across the entire region you're considering
Local minimum	A point at which f is smaller than any nearby points in a small neighborhood
Critical point	A point where $f'(x) = 0$

Table 4: 18.01 terminology for critical points

Each row includes all the ones above it, but not vice-versa. Here's a picture of an example showing these for a random function $f(x) = -\frac{1}{5}x^6 - \frac{2}{7}x^5 + \frac{2}{3}x^4 + x^3$. From left to right in Figure 16, there are four critical points:

- A local maximum (that isn't a global maximum), drawn in blue.
- A local minimum (that isn't a global minimum), draw in green.
- An critical inflection point — neither a local minimum *nor* a local maximum. Drawn in orange.
- A global maximum, drawn in purple.

Note there's no global minimum at all, since the function f goes to $-\infty$ in both directions as $x \to -\infty$ or $x \to +\infty$.

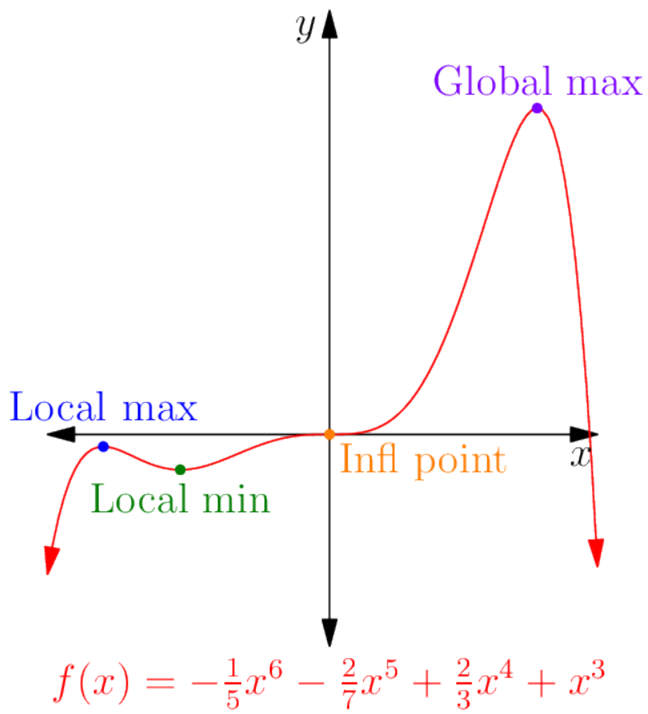

Figure 16: Some examples of critical points in an 18.01 graph of a single variable function.

§16.2 [TEXT] Critical points in 18.02

In 18.02, when we consider $f : \mathbb{R}^n \to \mathbb{R}$ the only change we make is:

Definition

For 18.02, we generalize the definition of **critical point** to be a point P for which $\nabla f(P) = \mathbf{0}$ is the zero vector. (The other two definitions don't change.)

As soon as I say this I need to carry over the analogous warnings from 18.01:

Warning

- Keep in mind that each of the implications

$$\text{Global minimum} \implies \text{Local minimum} \implies \text{Critical point, i.e. } \nabla f = \mathbf{0}$$

  is true only one way, not conversely. So a local minimum may not be a global minimum; and a point with gradient zero might not be a minimum, even locally. You should still find all the critical points, just be aware a lot of them may not actually be min's or max's.
- There may not be *any* global minimum or maximum at all, like we just saw.

Definition

In 18.02, a critical point that isn't a local minimum or maximum is called a **saddle point**.

Example

The best example of a saddle point to keep in your head is the origin for the function

$$f(x, y) = x^2 - y^2.$$

Why is this a saddle point? We have $f(0, 0) = 0$, and the gradient is zero too, since

$$\nabla f = \begin{pmatrix} 2x \\ 2y \end{pmatrix} \implies \nabla f(0, 0) = \begin{pmatrix} 2 \cdot 0 \\ 2 \cdot 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The problem is that the small changes in x and y clash in sign. Specifically, if we go a little bit to either the left or right in the x-direction, then f will *increase* a little bit, e.g.

$$f(0.1, 0) = f(-0.1, 0) = 0.01 > 0.$$

But the $-y^2$ term does the opposite: if we go a little bit up or down in the y-direction, then f will *decrease* a little bit.

$$f(0, 0.1) = f(0, -0.1) = -0.01 < 0.$$

So the issue is the clashing signs of small changes in x and y directions. This causes f to neither be a local minimum nor local maximum.

There's actually nothing special about $\pm x$ and $\pm y$ in particular; I only used those to make arithmetic easier. You can see Figure 17 for values of f at other nearby points.

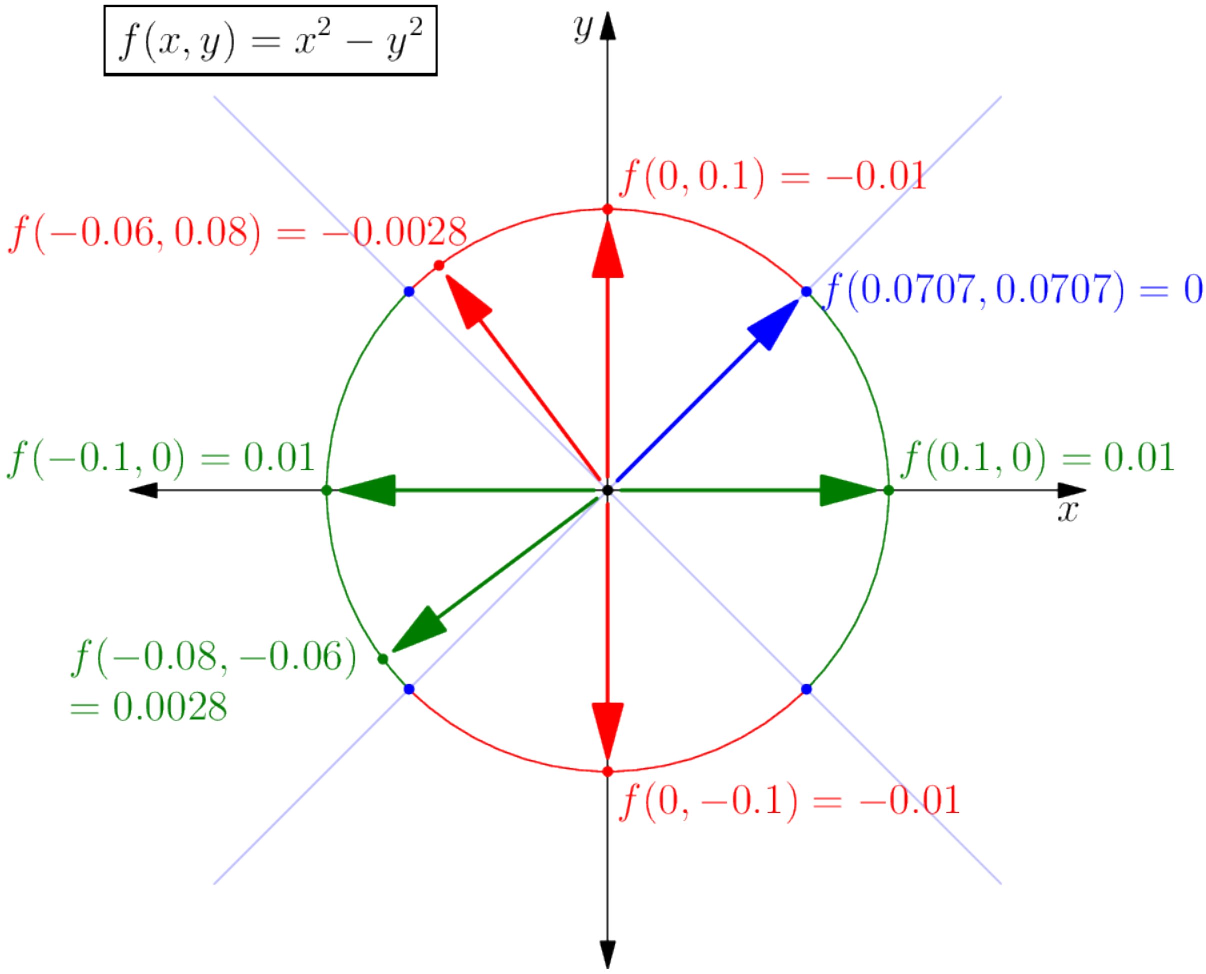

Figure 17: Values of $f(x, y) = x^2 - y^2$ at a distance of 0.1 from the saddle point $(0, 0)$. Green values are positive and red ones are negative. It's a saddle point because there are both.

Remark

The name "saddle point" comes from the following picture: if one looks at the surface

$$z = x^2 - y^2$$

then near $(0, 0)$ you have something that looks like a horse saddle. It curves upwards along the x-direction, but downwards along the y-direction.

We'll get to the recipe for distinguishing between saddle points and local minimums and maximums in a moment; like in 18.01, there is something called the second derivative test. First, one digression and a few examples of finding critical points.

§16.3 [SIDENOTE] Saddle points are way more common than critical inflection points

At first glance, you might be tempted to think that a saddle point is the 18.02 version of the critical inflection point. However, that analogy is actually not so good for your instincts, and **saddle points feel quite different from 18.01 critical inflection points**. Let me explain why.

In 18.01, it was *possible* for a critical point to be neither a local minimum or maximum, and we called these critical inflection points. However, in 18.01 this was actually really rare. To put this in perspective, suppose we considered a random 18.01 function of the form

$$f(x) = \square x^3 + \square x^2 + \square x + \square$$

where each square was a random integer between -1000000 and 1000000 inclusive. Of the approximately 10^{25} functions of this shape, you will find that while there are plenty of critical points, the chance of finding a critical inflection point is something like 10^{-15} — far worse than the lottery. (Of course, if you *know* where to look, you can find them: $f(x) = x^3$ has a critical inflection point at the origin, for example.)

In 18.02 this is no longer true. If we picked a random function of a similar form

$$f(x) = \square x^3 + \square x^2 + \square x + \square y^3 + \square y^2 + \square y + \square$$

where we fill each square with a number from -1000000 to 1000000 then you'll suddenly see saddle points everywhere. For example, when I ran this simulation 10000 times, among the critical points that showed up, I ended up with about

- 24.6% local minimums
- 25.3% local maximums
- 50.1% saddle points.

And the true limits (if one replaces 10^6 with N and takes the limit as $N \to \infty$) are what you would guess from the above: 25%, 25%, 50%. (If you want to see the code, it's in the Appendix, Section 27.3.)

Why is the 18.02 situation so different? It comes down to this: in 18.02, you can have two clashing directions. For the two experiments I've run here, consider the picture in Figure 18. Here P is a critical point, and we consider walking away from it in one of two directions. I'll draw a blue $+$ arrow if f increases, and a red $-$ arrow if f decreases.

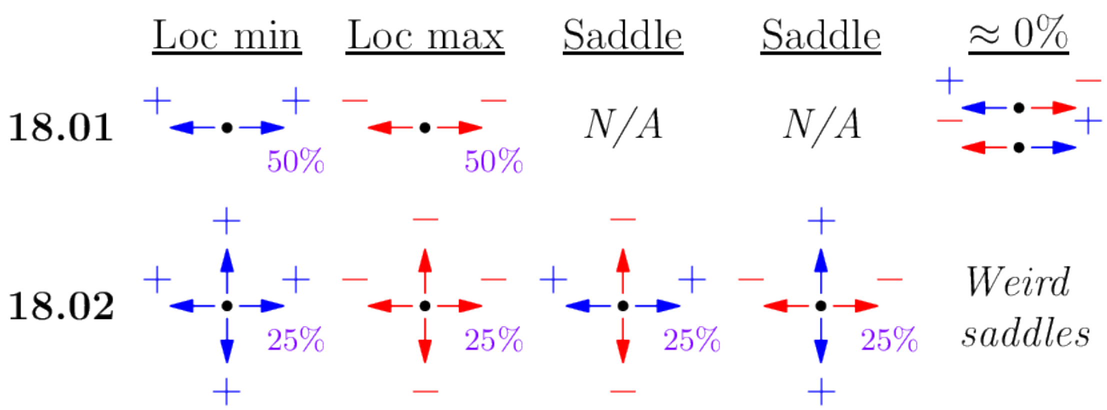

Figure 18: Why the 18.01 and 18.02 polynomial experiments have totally different outcomes.

In the 18.01 experiment, we saw that two arrows pointing *opposite* directions almost always have the same color. So in 18.01, when we could only walk in one direction, that meant almost every point was either a local minimum or a local maximum. But the picture for 18.02 is totally different because there's nothing that forces the north/south pair to have the same sign as the east/west pair. For a "random" function, if you believe the colors are equally likely, then half the time the arrows don't match colors and you end up with a saddle point.

This whole section was for two-variable functions $P(x) + Q(y)$, so it's already a simplification. If you ran an analogous three-variable experiment defined similarly for polynomials $f(x, y, z) = P(x) + Q(y) + R(z)$:

- 12.5% local minimums
- 12.5% local maximums
- 75.0% saddle points.

If we return to the world of *any* two-variable function, the truth is even more complicated than this. In this sidenote I only talked about functions $f(x, y)$ that looked like $P(x) + Q(y)$ where P and Q were polynomials. The x and y parts of the function were completely unconnected, so we only looked in the four directions north/south/east/west. But most two-variable functions have some more dependence between x and y, like $f(x, y) = x^2y^3$ or $f(x) = e^x \sin(y)$ or similar. Then you actually need to think about more directions than just north/south/east/west.

> **Digression**
>
> For example, Poonen's lecture notes (see question 9.22) show a weird *monkey saddle*: the point $(0, 0)$ is a critical point of
>
> $$f(x, y) = xy(x - y)$$
>
> where the values of f nearby split into six regions, alternating negative and positive, in contrast to Figure 17 where there were only four zones on the circle. (See also Wikipedia for monkey saddle.) Poonen also invites the reader to come up with an *octopus saddle* (which sounds like it needs sixteen regions, eight down ones for each leg of the octopus).

§16.4 [RECIPE] Finding critical points

For finding critical points, on paper you can just follow the definition:

> **Recipe for finding critical points**
>
> To find the critical points of $f : \mathbb{R}^n \to \mathbb{R}$
>
> 1. Compute the gradient ∇f.
> 2. Set it equal to the zero vector and solve the resulting system of n equations in n variables.

The thing that might be tricky is that you have to solve a system of equations. Depending on how difficult your function is to work with, that might require some creativity in order to get the algebra right. We'll show some examples where the algebra is really simple, and examples where the algebra is much more involved.

> **Sample Question**
>
> Find the critical points of $f(x, y, z) = x^2 + 2y^2 + 3z^2$.

Solution: The gradient is

$$\nabla f(x, y, z) = \begin{pmatrix} 2x \\ 4y \\ 6z \end{pmatrix}.$$

In order for this to equal $\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$, we need to solve the three-variable system of equations

$$\begin{aligned} 2x &= 0 \\ 4y &= 0 \\ 6z &= 0 \end{aligned}$$

which is so easy that it's almost insulting: $x = y = z = 0$. The only critical point is $(0, 0, 0)$. $\square$

Sample Question

Find the critical points of $f(x, y) = xy(6 - x - y)$.

Solution: This example is a lot more annoying than the previous one, despite having fewer variables, because casework is forced upon you. You need to solve four systems of linear equations, not just one, as you'll see.

We expand

$$f(x, y) = 6xy - x^2y - xy^2.$$

So

$$\nabla f = \begin{pmatrix} 6y - 2xy - y^2 \\ 6x - x^2 - 2xy \end{pmatrix}.$$

Hence, the resulting system of equations to solve is

$$\begin{aligned} y(6 - 2x - y) &= 0 \\ x(6 - 2y - x) &= 0. \end{aligned}$$

The bad news is that these are quadratic equations. Fortunately, they come in factored form, so we can rewrite them as OR statements:

$$\begin{aligned} y(6 - 2x - y) = 0 &\implies (y = 0 \quad \text{OR } 2x + y = 6) \\ x(6 - 2y - x) = 0 &\implies (x = 0 \quad \text{OR } x + 2y = 6). \end{aligned}$$

So actually there are $2^2 = 4$ cases to consider, and we have to manually tackle all four. These cases fit into the following 2×2 table; we solve all four systems of equations.

	Top eqn. gives $y = 0$	Top eqn. gives $2x + y = 6$
Bottom eqn. gives $x = 0$	$\begin{cases} y=0 \\ x=0 \end{cases} \implies (x, y) = (0, 0)$	$\begin{cases} 2x+y=6 \\ x=0 \end{cases} \implies (x, y) = (0, 6)$
Bottom eqn. gives $x + 2y = 6$	$\begin{cases} y=0 \\ x+2y=6 \end{cases} \implies (x, y) = (6, 0)$	$\begin{cases} 2x+y=6 \\ x+2y=6 \end{cases} \implies (x, y) = (2, 2)$

,

So we get there are four critical points, one for each case: $(0, 0)$, $(0, 6)$, $(6, 0)$ and $(2, 2)$. $\square$

§16.5 [RECIPE] The second derivative test for two-variable functions

Earlier we classified critical points by looking at nearby points. Technically speaking, we did not give a precise definition of "nearby", just using small numbers like 0.01 or 0.1 to make a point. So in 18.02, the exam will want a more systematic theorem for classifying critical points as local minimum, local maximum, or saddle point.

I thought for a bit about trying to explain why the second derivative test works, but ultimately I decided to not include it in these notes. Here's some excuses why:

> **Digression**
>
> The issue is that getting the "right" understanding of this would require me to talk about *quadratic forms*. However, in the prerequisite parts Alfa and Bravo of these notes, we only did linear algebra, and didn't cover quadratic forms in this context at all. I hesitate to introduce an entire chapter on quadratic forms (which are much less intuitive than linear functions) and *then* tie that to eigenvalues of a 2×2 matrix just to justify a single result that most students will just memorize anyway.
>
> Poonen has some hints on quadratic forms in section 9 of his notes if you want to look there though.
>
> The other downside is that even if quadratic forms are done correctly, the second derivative test doesn't work in all cases anyway, if the changes of the function near the critical point are subquadratic (e.g. degree three). And multivariable Taylor series are not on-syllabus for 18.02.

So to get this section over with quickly, I'll just give the result. I'm sorry this will seem to come out of nowhere.

> **Recipe: The second derivative test**
>
> Suppose $f(x, y)$ is a function has a critical point at P. We want to tell whether it's a local minimum, local maximum, or saddle point. Assume that f has a continuous second derivative near P.
>
> 1. Let $A = f_{xx}(P)$, $B = f_{xy}(P) = f_{yx}(P)$, $C = f_{yy}(P)$. These are the partial derivatives of the partial derivatives of f (yes, I'm sorry), evaluated at P. If you prefer gradients, you could write this instead as
>
> $$\nabla f_x(P) = \begin{pmatrix} A \\ B \end{pmatrix}, \quad \nabla f_y(P) = \begin{pmatrix} B \\ C \end{pmatrix}.$$
>
> 2. If $AC - B^2 \neq 0$, output the answer based on the following chart:
>   - If $AC - B^2 < 0$, output "saddle point".
>   - If $AC - B^2 > 0$ and $A, C \geq 0$, output "local minimum".
>   - If $AC - B^2 > 0$ and $A, C \leq 0$, output "local maximum".
> 3. If $AC - B^2 = 0$, the second derivative test is inconclusive. Any of the above answers are possible, including weird/rare saddle points like the monkey saddle. You have to use a different method instead.

Tip

It is indeed a theorem that if f is differentiable twice continuously, then $f_{xy} = f_{yx}$. That is, if you take a well-behaved function f and differentiate with respect to x then differentiate with respect to y, you get the same answer as if you differentiate with respect to y and respect to x. You'll see this in the literature written sometimes as

$$\frac{\partial}{\partial x}\frac{\partial}{\partial y}f = \frac{\partial}{\partial y}\frac{\partial}{\partial x}f.$$

Sample Question

Use the second derivative test to classify the critical point $(0,0)$ of the function

$$f(x,y) = x^3 + x^2 + y^3 - y^2.$$

Solution: Start by computing the partial derivatives:

$$\nabla f = \begin{pmatrix} 3x^2 + 2x \\ 3y^2 - 2y \end{pmatrix} \implies \begin{cases} f_x = 3x^2 + 2x \\ f_y = 3y^2 - 2y \end{cases}.$$

We now do partial differentiation a second time on each of these. Depending on your notation, you can write this as either

$$\nabla f_x = \begin{pmatrix} 6x + 2 \\ 0 \end{pmatrix} \quad \nabla f_y = \begin{pmatrix} 0 \\ 6y - 2 \end{pmatrix}$$

or

$$f_{xx} = 6x + 2, \quad f_{xy} = f_{yx} = 0, \quad f_{yy} = 6y - 2.$$

Again, the repeated $f_{xy} = f_{yx}$ is either $\frac{\partial}{\partial y}(6x+2) = 0$ or $\frac{\partial}{\partial x}(6y-2) = 0$; for well-behaved functions, you always get the same answer for f_{xy} and f_{yx}.

At the origin, we get

$$\begin{aligned} A &= 6 \cdot 0 + 2 = 2 \\ B &= 0 \\ C &= 6 \cdot 0 - 2 = -2. \end{aligned}$$

Since $AC - B^2 = -4 < 0$, we output the answer "saddle point". □

Sample Question

Find the critical points of $f(x,y) = xy + y^2 + 2y$ and classify them using the second derivative test.

Solution: Start by computing the gradient:

$$\nabla f = \begin{pmatrix} y \\ x + 2y + 2 \end{pmatrix}.$$

Solve the system of equations $y = 0$ and $x + 2y + 2 = 0$ to get just $(x, y) = (-2, 0)$. Hence this is the only critical point.

We now compute the second derivatives:

$$f_{xx} = \frac{\partial}{\partial x}(y) = 0$$
$$f_{xy} = f_{yx} = \frac{\partial}{\partial y}(y) = \frac{\partial}{\partial x}(x + 2y + 2) = 1$$
$$f_{yy} = \frac{\partial}{\partial y}(x + 2y + 2) = 2.$$

These are all constant functions in this example; anyway, we have $A = 0$, $B = 1$, $C = 2$, and $AC - B^2 = -1 < 0$, so output "saddle point". □

§16.6 [EXER] Exercises

*** Exercise 16.1**:

- Give an example of a differentiable function $f : \mathbb{R}^2 \to \mathbb{R}$ with the following property: every lattice point (x, y) (i.e. a point where both x and y are integers) is a saddle point, and there are no other saddle points.
- Does there exist a differentiable function $f : \mathbb{R}^2 \to \mathbb{R}$ such that every point is a saddle point?

§17 Regions

In 18.02, you'll be asked to find global minimums or maximums over a **constraint region** $\mathcal{R}$, which is only a subregion of $\mathbb{R}^n$. For example, if you have a three-variable function $f(x, y, z)$ given to you, you may be asked questions like

- What is the global maximum of f (if any) across all of $\mathbb{R}^3$?
- What is the global maximum of f (if any) across the octant[10] $x, y, z > 0$?
- What is the global maximum of f (if any) across the cube given by $-1 \leq x, y, z \leq 1$?
- What is the global maximum of f (if any) across the sphere $x^2 + y^2 + z^2 = 1$?
- ... and so on.

It turns out that thinking about constraint regions is actually half the problem. In 18.01 you usually didn't have to think much about it, because the regions you got were always intervals, and that made things easy. But in 18.02, you will need to pay much more attention.

> **Warning: if you are proof-capable, read the grown-up version**
>
> This entire section is going to be a lot of wishy-washy terms that I don't actually give definitions for. If you are a high-school student preparing for a math olympiad, or you are someone who can read proofs, read the version at `https://web.evanchen.cc/handouts/LM/LM.pdf` instead. We use open/closed sets and compactness there to do things correctly.

§17.1 [TEXT] Constraint regions

> **Digression: An English lesson on circle vs disk, sphere vs ball**
>
> To be careful about some words that are confused in English, I will use the following conventions:
>
> - The word **circle** refers to a one-dimensional object with no inside, like $x^2 + y^2 = 1$. It has no area.
> - The word **open disk** refers to points strictly inside a circle, like $x^2 + y^2 < 1$
> - The word **closed disk** refers to a circle and all the points inside it, like $x^2 + y^2 = 1$ or $x^2 + y^2 < 1$.
> - The word **disk** refers to either an open disk or a closed disk.
>
> Similarly, a **sphere** refers only to the surface, not the volume, like $x^2 + y^2 + z^2 = 1$. Then we have **open ball**, **closed ball**, and **ball** defined in the analogous way.

In 18.02, all the constraint regions we encounter will be made out of some number (possibly zero) of equalities and inequalities. We provide several examples.

> **Examples of regions in $\mathbb{R}$**
>
> In $\mathbb{R}$:
> - All of $\mathbb{R}$, with no further conditions.
> - An open interval like $-1 < x < 1$ in $\mathbb{R}$.
> - A closed interval like $-1 \leq x \leq 1$ in $\mathbb{R}$.

[10] Like "quadrant" with xy-graphs. If you've never seen this word before, ignore it.

Examples of two-dimensional regions in $\mathbb{R}^2$

In $\mathbb{R}^2$, some two-dimensional regions:

- All of $\mathbb{R}^2$, with no further conditions.
- The first quadrant $x, y > 0$, not including the axes
- The first quadrant $x, y \geq 0$, including the positive x and y axes.
- The square $-1 < x < 1$ and $-1 < y < 1$, not including the four sides of the square.
- The square $-1 \leq x \leq 1$ and $-1 \leq y \leq 1$, including the four sides.
- The open disk $x^2 + y^2 < 1$, filled-in unit disk without its circumference.
- The closed disk $x^2 + y^2 \leq 1$, filled-in unit disk including its circumference.

Examples of one-dimensional regions in $\mathbb{R}^2$

In $\mathbb{R}^2$, some one-dimensional regions:

- The unit circle $x^2 + y^2 = 1$, which is a circle of radius 1, not filled.
- Both $x^2 + y^2 = 1$ and $x, y > 0$, a quarter-arc, not including $(1, 0)$ and $(0, 1)$.
- Both $x^2 + y^2 = 1$ and $x, y \geq 0$, a quarter-arc, including $(1, 0)$ and $(0, 1)$.
- The equation $x + y = 1$ is a line.
- Both $x + y = 1$ and $x, y > 0$: a line segment not containing the endpoints $(1, 0)$ and $(0, 1)$.
- Both $x + y = 1$ and $x, y \geq 0$: a line segment containing the endpoints $(1, 0)$ and $(0, 1)$.

I could have generated plenty more examples for $\mathbb{R}^2$, and I haven't even gotten to $\mathbb{R}^3$ yet. That's why the situation of constraint regions requires more thought in 18.02 than 18.01, (whereas in 18.01 there were pretty much only a few examples that happened).

In order to talk about the regions further, I have to introduce some new words. The three that you should care about for this class are the following: "boundary", "limit cases", and "dimension".

⚠ Warning

As far as I know, in 18.02 it's not possible to give precise definitions for these words. So you have to play it by ear. All the items below are rules of thumb that work okay for 18.02, but won't hold up in 18.100/18.900.

- The **boundary** is usually the points you get when you choose any one of the $\leq$ and $\geq$ constraints and turn it into and $=$ constraint. For example, the boundary of the region cut out by $-1 \leq x \leq 1$ and $-1 \leq y \leq 1$ (which is a square of side length 2) are the four sides of the square, where either $x = \pm 1$ or $y = \pm 1$.
- The **limit cases** come in two forms:
  - If any of the variables can go to $\pm\infty$, all those cases are usually limit cases.
  - If you have any $<$ and $>$ inequalities, the cases where the variables approach those strict bounds are usually limit cases.
- The **dimension** of $\mathcal{R}$ is the hardest to define in words but easiest to guess. I'll give you two ways to guess it:
  - Geometric guess: pick a point P in $\mathcal{R}$ that's not on the boundary. Look at all the points of $\mathcal{R}$ that are close to P, i.e. a small neighborhood.

  - Say $\mathcal{R}$ is one-dimensional if the small neighborhood could be given a *length*.
  - Say $\mathcal{R}$ is two-dimensional if the small neighborhood could be given an *area*.
  - Say $\mathcal{R}$ is three-dimensional if the small neighborhood could be given a *volume*.
- Algebraic guess: the dimension of a region in $\mathbb{R}^n$ is usually equal to n minus the number of $=$ in constraints.

Overall, trust your instinct on dimension; you'll usually be right.

The table below summarizes how each constraint affects each of the three words above.

Constraint	Boundary	Limit case	Dimension
$\le$ or $\ge$	Change to $=$ to get boundary	No effect	No effect
$<$ or $>$	No effect	Approach for limit case	No effect
$=$	No effect	No effect	Reduces dim by one

Table 5: Effects of the rules of thumb.

Let's use some examples.

Example: the circle, open disk, and closed disk

- The circle $x^2 + y^2 = 1$ is a **one-dimensional** shape. Again, we consider this region to be *one-dimensional* even though the points live in $\mathbb{R}^2$. The rule of thumb is that with 2 variables and 1 equality, the dimension should be $2 - 1 = 1$.

  Because there are no inequality constraints at all, and because x and y can't be larger than 1 in absolute value, there is no **boundary** and there are no **limit cases**.
- The open disk $x^2 + y^2 < 1$ is **two-dimensional** now, since it's something that makes sense to assign an area. (Or the rule of thumb that with 2 variables and 0 equalities, the dimension should be $2 - 0 = 2$.)

  There is one family of **limit cases: when** $x^2 + y^2$ **approaches** 1^-. But there is no boundary.
- The closed disk $x^2 + y^2 \le 1$ is also **two-dimensional**. Because x and y can't be larger than 1 in absolute value, and there were no $<$ or $>$ constraints, there are no limit cases to consider. But there is a **boundary of** $x^2 + y^2 = 1$.

TODO: Draw a figure for this

In compensation for the fact that I'm not giving you true definitions, I will instead give you a pile of examples, their dimensions, boundaries, and limit cases. See Table 6, Table 7, Table 8.

Region	Dim.	Boundary	Limit cases
All of $\mathbb{R}$	1-D	No boundary	$x \to \pm\infty$
$-1 < x < 1$	1-D	No boundary	$x \to \pm 1$
$-1 \le x \le 1$	1-D	$x = \pm 1$	No limit cases

Table 6: Examples of regions inside $\mathbb{R}$ and their properties.

Region	**Dim.**	**Boundary**	**Limit cases**
All of $\mathbb{R}^2$	2-D	No boundary	$x \to \pm\infty$ or $y \to \pm\infty$
$x, y > 0$	2-D	No boundary	$x \to 0^+$ or $y \to 0^+$ or $x \to +\infty$ or $y \to +\infty$
$x, y \geq 0$	2-D	$x = 0$ or $y = 0$	$x \to +\infty$ or $y \to +\infty$
$-1 < x < 1$ $-1 < y < 1$	2-D	No boundary	$x, y \to \pm 1$
$-1 \leq x \leq 1$ $-1 \leq y \leq 1$	2-D	$x = \pm 1$ or $y = \pm 1$	No limit cases
$x^2 + y^2 < 1$	2-D	No boundary	$x^2 + y^2 \to 1^-$
$x^2 + y^2 \leq 1$	2-D	$x^2 + y^2 = 1$	No limit cases
$x^2 + y^2 = 1$	1-D	No boundary	No limit cases
$x^2 + y^2 = 1$ $x, y > 0$	1-D	No boundary	$x \to 0^+$ or $y \to 0^+$
$x^2 + y^2 = 1$ $x, y \geq 0$	1-D	$(1, 0)$ and $(0, 1)$	No limit cases
$x + y = 1$	1-D	No boundary	$x \to \pm\infty$ or $y \to \pm\infty$
$x + y = 1$ $x, y > 0$	1-D	No boundary	$x \to 0^+$ or $y \to 0^+$
$x + y = 1$ $x, y \geq 0$	1-D	$(1, 0)$ and $(0, 1)$	No limit cases

Table 7: Examples of regions inside $\mathbb{R}^2$ and their properties

Region	**Dim.**	**Boundary**	**Limit cases**
All of $\mathbb{R}^3$	3-D	No boundary	Any var to $\pm\infty$
$x, y, z > 0$	3-D	No boundary	Any var to 0 or ∞
$x, y, z \geq 0$	3-D	$x = 0$ or $y = 0$ or $z = 0$	Any var to ∞
$x^2 + y^2 + z^2 < 1$	3-D	No boundary	$x^2 + y^2 + z^2 \to 1^-$
$x^2 + y^2 + z^2 \leq 1$	3-D	$x^2 + y^2 + z^2 = 1$	No limit cases
$x^2 + y^2 + z^2 = 1$	2-D	No boundary	No limit cases
$x^2 + y^2 + z^2 = 1$ $x, y, z > 0$	2-D	No boundary	$(1, 0)$ and $(0, 1)$
$x^2 + y^2 + z^2 = 1$ $x, y, z \geq 0$	2-D	Three quarter-circle arcs[11]	No limit cases
$x + y + z = 1$	2-D	No boundary	Any var to $\pm\infty$
$x + y + z = 1$ $x, y, z > 0$	2-D	No boundary	Any var to 0^+
$x + y + z = 1$ $x, y, z \geq 0$	2-D	$x = 0$ or $y = 0$ or $z = 0$	No limit cases

Table 8: Examples of regions inside $\mathbb{R}^3$ and their properties

[11]To be explicit, the first quarter circle is $x^2 + y^2 = 1, x, y \geq 0$ and $z = 0$. The other two quarter-circle arcs are similar.

Digression on intentionally misleading constraints that break the rule of thumb

I hesitate to show these, but here are some examples where the rules of thumb fail:

- An unusually cruel exam-writer might rewrite the unit circle as
$$x^2 + y^2 \leq 1 \ \text{ and } x^2 + y^2 \geq 1$$
instead of the more natural $x^2 + y^2 = 1$. Then if you were blindly following the rules of thumb, you'd get the wrong answer.
- In $\mathbb{R}^3$ the region cut out by the single equation
$$x^2 + y^2 + z^2 = 0$$
is actually 0-dimensional, because there's only one point in it: $(0, 0, 0)$.

That said, intentionally misleading constraints like this are likely off-syllabus for 18.02.

§17.2 [RECIPE] Working with regions

This is going to be an unsatisfying recipe, because it's just the rules of thumb. But again, for 18.02, the rules of thumb should work on all the exam questions.

Recipe: The rule of thumb for regions

Given a region $\mathcal{R}$ contained in $\mathbb{R}^n$, to guess its dimension, limit cases, and boundary:

- The dimension is probably n minus the number of $=$ constraints.
- The limit cases are obtained by turning $<$ and $>$ into limits, and considering when any of the variables can go to $\pm\infty$.
- The boundary is obtained when any $\leq$ and $\geq$ becomes $=$.

See Table 5.

TODO: Add some more examples here

§18 Optimization problems

Now that we understand both critical points of f and regions $\mathcal{R}$, we turn our attention to the problem of finding global minimums and maximums.

§18.1 [TEXT] The easy and hard cases

Suppose you have a function $f : \mathbb{R}^n \to \mathbb{R}$ that you can compute ∇f for, and a region $\mathcal{R}$. We're going to distinguish between two cases:

- The **easy case** is if $\mathcal{R}$ has dimension n as well. The rule of thumb says there should be zero "=" constraints.
- The **hard case** is if $\mathcal{R}$ has dimension $n - 1$. Rule of thumb says there should be one "=" constraint. In the hard case, we will use **Lagrange multipliers**.

We won't cover the case where $\mathcal{R}$ has dimension $n - 2$ or less in 18.02 (i.e. two or more constraints), although it can be done.

§18.2 [RECIPE] The easy case

Recipe for optimization without Lagrange Multipliers

Suppose you want to find the optimal values of $f : \mathbb{R}^n \to \mathbb{R}$ over a region $\mathcal{R}$, and $\mathcal{R}$ has dimension n.

1. Evaluate f on all the **critical points** of f in the region $\mathcal{R}$.
2. Evaluate f on all the **boundary points** of f in the region $\mathcal{R}$.
3. Evaluate f on all the **limit cases** of f in the region $\mathcal{R}$.
4. Output the points in the previous steps that give the best values, or assert the optimal value doesn't exist (if points in step 3 do better than steps 1-2).

If there are any points at which ∇f is undefined, you should check those as well. However, these seem to be pretty rare for the examples that show up in 18.02.

Warning

Step 2 might actually require Lagrange multipliers, even in the easy case. Don't underestimate the difficulty of the boundary cases.

Sample Question

Find the minimum and maximum possible value, if they exist of

$$f(x, y) = x + y + \frac{8}{xy}$$

over $x, y > 0$.

Solution: The region $\mathcal{R}$ is the first quadrant which is indeed two-dimensional (no $=$ constraints), so we're in the easy case and the recipe applies here. We check all the points in turn:

1. To find the critical points, calculate the gradient

$$\nabla f(x, y) = \begin{pmatrix} 1 - \frac{8}{x^2 y} \\ 1 - \frac{8}{x y^2} \end{pmatrix}$$

and then set it equal to $\binom{0}{0}$. This gives us the simultaneous equations

$$1 = \frac{8}{x^2 y} = \frac{8}{x y^2}.$$

This implies $x^2 y = x y^2$ or $x = y$ (we have $x, y > 0$ in $\mathcal{R}$, so we're not worried about division by zero) and so the only critical point is $(x, y) = (2, 2)$.

2. The region $\mathcal{R}$ has no boundary, so there are no boundary points to check.
3. The region $\mathcal{R}$ has four different kinds of limit cases:
   - $x \to 0^+$
   - $y \to 0^+$
   - $x \to +\infty$
   - $y \to +\infty$.

   In fact all four of these cases cause $f \to +\infty$. In each of the first two cases, the term $\frac{8}{xy}$ in f causes $f \to +\infty$. In the case $x \to \infty$, the term x causes $f \to +\infty$. In the case $y \to \infty$, the term y causes $f \to +\infty$.

Putting these together:

- The global minimum is $(2, 2)$, at which $f(2, 2) = 6$.
- There is no global maximum, since we saw limit cases where $f \to +\infty$. □

§18.3 [TEXT] Lagrange multipliers

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function we're optimizing over some region $\mathcal{R}$. We now turn to the case where $\mathcal{R}$, is dimension $n - 1$, because of a single constraint of the form $g(x, y) = c$ or $g(x, y, z) = c$.

We need a new definition of critical point. To motivate it, let's consider a particular example in Figure 19. Here $n = 2$, and

- $f(x, y) = x^2 + y^2$, and
- $g(x, y) = c$ is the red level curve shown in the picture below;
- $\mathcal{R}$ is just the level curve $g(x, y) = c$ (no further $<$ or $\leq$ constraints).

Trying to optimize f subject to $g(x, y) = c$ in this picture is the same as finding the points on the level curve which are furthest or closest to the origin. I've marked those two points as P and Q in the figure. The trick to understanding how to get them is to *also* draw the level curves for f that pass through P and Q: then we observe that the level curves for f that get those minimums and maximums ought to be tangent to $g(x, y) = c$ at P and Q.

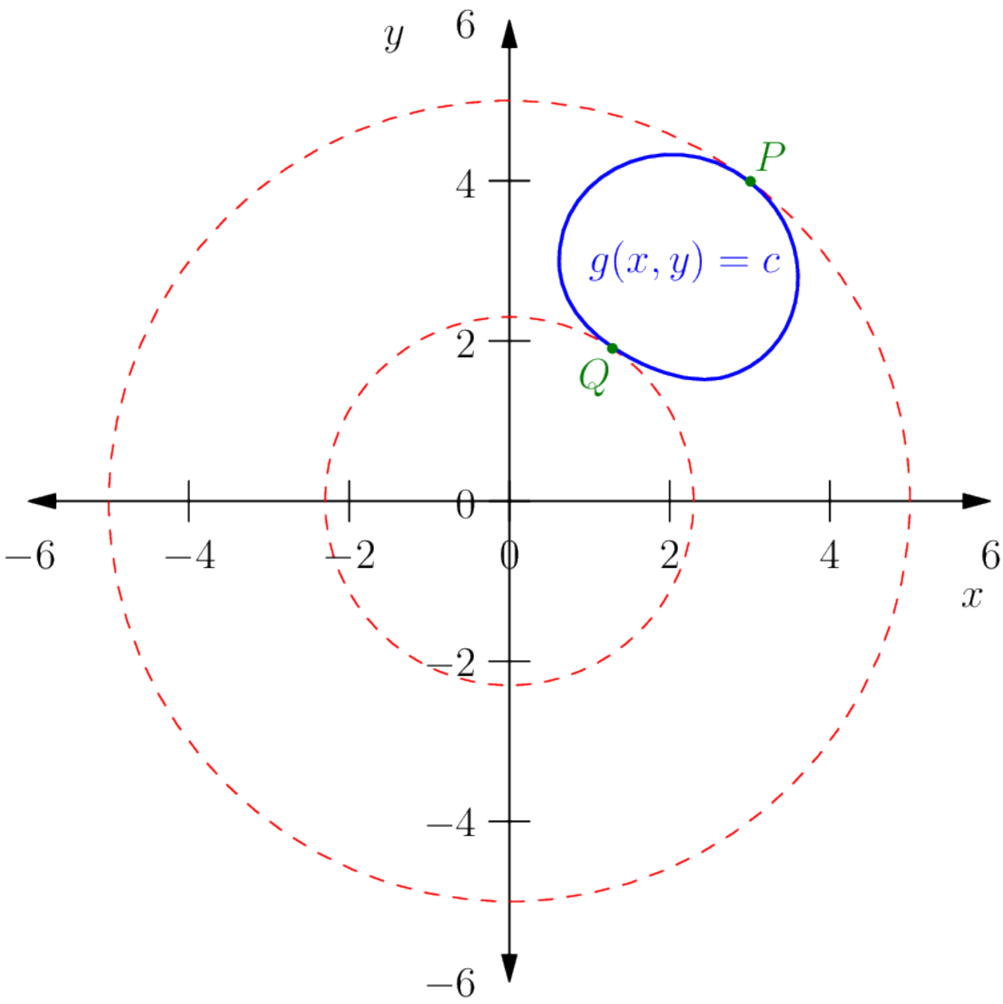

Figure 19: An example of a LM-type optimization problem, where one finds points on $g(x, y) = c$ which optimize f

Now how can we check whether there's a tangency? Answer: look at the gradient! We expect that ∇f and ∇g, at the points P and Q, should point in the same direction. So that gives us the strategy: look for the points where ∇f and ∇g point the same way.

I don't think the following term is an official name, but I like it, and I'll use it:

Definition

An **LM-critical point** is a point P such that either

- $\nabla f(P) = \lambda \nabla g(P)$ for some scalar λ; or
- $\nabla g(P) = 0$.

Note that there are *two* hypotheses. If you want, you can think about this as requiring that $\nabla f(P)$ and $\nabla g(P)$ are *linearly dependent*, so it's only one item. However, in practice, people end up usually breaking into cases like this.

> **Digression**
>
> The parameter λ is the reason for the name "Lagrange multipliers"; it's a scalar multiplier on ∇g. Personally, I don't think this name makes much sense.

Now that we have this, we can describe the recipe for the "hard" case. The only change is to replace the old critical point definition (where $\nabla f(P) = 0$) with the LM-critical point definition.

§18.4 [RECIPE] Lagrange multipliers

> **Recipe for Lagrange multipliers**
>
> Suppose you want to find the optimal values of $f : \mathbb{R}^n \to \mathbb{R}$ over a region $\mathcal{R}$, and $\mathcal{R}$ has dimension $n-1$ due to a single constraint $g = c$ for some $g : \mathbb{R}^n \to \mathbb{R}$.
>
> 1. Evaluate f on all the **LM-critical points** of f that lie on the region $\mathcal{R}$.
> 2. Evaluate f on all the **boundary points** of f of the region $\mathcal{R}$.
> 3. Evaluate f on all the **limit cases** of f of the region $\mathcal{R}$.
> 4. Output the points in the previous steps that give the best values, or assert the optimal value doesn't exist (if points in step 3 do better than steps 1-2).
>
> If there are any points at which ∇f or ∇g are undefined, you should check those as well. However, these seem to be pretty rare for the examples that show up in 18.02.

Again, this is the same recipe as Section 18.2, except we changed "critical point" to "LM-critical point".

> **Tip**
>
> Remember how finding critical points could lead to systems of equations that required quite a bit of algebraic skill to solve? The same is true for Lagrange multipliers, but even more so, because of the new parameter λ that you have to care about. So the reason this is called the "hard case" isn't because the 18.02 ideas needed are different, but because the algebra can become quite involved in finding LM-critical points.
>
> In fact, in high school math competitions, the algebra can sometimes become so ugly that the method of Lagrange multipliers is sometimes jokingly called "Lagrange *murderpliers*" to reflect the extreme amount of calculation needed for some problems.

> **Tip**
>
> When solving the system of equations, one strategy is to start by eliminating λ, because we don't usually care about the value of λ.

> **Sample Question**
>
> Find the minimum and maximum possible value, if they exist, of
>
> $$f(x, y, z) = x + y + z$$
>
> over $x, y, z > 0$ satisfying the condition $xyz = 8$.

Solution: The region $\mathcal{R}$ is two-dimensional, consisting of strict inequalities $x, y, z > 0$ and the condition $g(x, y, z) = xyz = 8$. We carry out the recipe.

1. To find the LM-critical points, we need to compute both ∇f and ∇g. We do so:
$$\nabla f(x, y, z) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$
$$\nabla g = (yz, zx, xy).$$
Now, there are no points with $\nabla g = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ in the region $\mathcal{R}$, because in $\mathcal{R}$ all the variables are constrained to be positive. So we now solve the system
$$\begin{aligned} 1 &= \lambda yz \\ 1 &= \lambda zx \\ 1 &= \lambda xy \end{aligned}$$
and see what values it takes.

   The trick to solving the system of equations is to divide the first two to get rid of the parameter λ, which we don't really care about, to get
$$\frac{1}{1} = \frac{\lambda yz}{\lambda zx} = \frac{y}{x}.$$
So we must have $x = y$. Similarly, we find $y = z$ and $z = x$.

   Hence the LM-critical point must have $x = y = z$. Since $xyz = 8$, it follows the only LM-critical point is $(2, 2, 2)$. Evaluating f here gives $f(2, 2, 2) = 6$.
2. The region $\mathcal{R}$ has no boundary, because no $\leq$ or $\geq$ constraints are present.
3. The region $\mathcal{R}$ has limit cases when any of the variables x, y, z either approach 0 or $+\infty$. However, remember that $xyz = 8$. So if any variable approaches 0, some other variable must become large. Consequently, in every limit case, we find that $f \to +\infty$.

Collating all these results:

- The unique global minimum is $(2, 2, 2)$ at which $f(2, 2, 2) = 6$.
- There is no global maximum. □

Remark

If you're paying close enough attention, you might realize this sample question we just did is a thin rewriting of the example in Section 18.2. This illustrates something: sometimes you can rewrite a hard-case optimization problem in 3 variables to an easy-case one with 2 variables.

The following sample question shows an optimization within an optimization problem. If you've seen the movie *Inception*, yes, one of those.

Sample Question

Find the minimum and maximum possible value, if they exist, of

$$f(x, y, z) = x^4 + y^4 + z^4$$

over the region $x^2 + y^2 + z^2 \le 1$.

Solution: At first glance, this seems like it should be in the easy case! The region $\mathcal{R}$ consisting of the closed ball $x^2 + y^2 + z^2 \le 1$ is indeed three-dimensional. But the reason this sample question is in this section is because we will find that checking the boundary case requires another application of Lagrange multipliers.

Let's carry out the easy case recipe.

1. First let's find the critical points of $f(x, y, z) = x^4 + y^4 + z^4$. Write

$$\nabla f = \begin{pmatrix} 4x^3 \\ 4y^3 \\ 4z^3 \end{pmatrix}.$$

   Solving the insultingly easy system of equations $4x^3 = 4y^3 = 4z^3 = 0$ we see the only critical point is apparently $x = y = z = 0$. The value there is $f(0, 0, 0) = 0$.

2. The boundary of $\mathcal{R}$ is $x^2 + y^2 + z^2 = 1$, the unit sphere; we denote this sphere by $\mathcal{S}$. So now we have to evaluate f on this boundary. The issue is that there are a lot of points on this unit sphere! We can't just check them one by one and there is on easy algebraic way to deal with them. Therefore, we will use Lagrange multipliers with the constraint function $g(x, y, z) = x^2 + y^2 + z^2$.

   1. Let's find the LM-critical points for f on $\mathcal{S}$. Take the gradient of g to get

$$\nabla g = \begin{pmatrix} 2x \\ 2y \\ 2z \end{pmatrix}.$$

      The only point at which $\nabla g = \mathbf{0}$ is $x = y = z = 0$ which isn't on the sphere $\mathcal{S}$, so we don't have to worry about $\nabla g = \mathbf{0}$ the case. Now we instead solve

$$\begin{pmatrix} 4x^3 \\ 4y^3 \\ 4z^3 \end{pmatrix} = \lambda \begin{pmatrix} 2x \\ 2y \\ 2z \end{pmatrix}.$$

      This requires some manual labor to solve, because there are lots of cases. The equation for x says that

$$4x^3 = \lambda \cdot 2x \Longleftrightarrow x = 0 \quad \text{or } x = \pm\sqrt{\frac{\lambda}{2}}$$

      and similarly for y and z.

      In other words, **all the nonzero variables** should have the same absolute value. (Think about why this sentence is true.) So if all three variables are nonzero, then $|x| = |y| = |z| = \frac{1}{\sqrt{3}}$ (because $x^2 + y^2 + z^2 = 1$ as well). If two variables are nonzero, then their absolute values are $\frac{1}{\sqrt{2}}$ by the same token. And if only one variable is nonzero, it is ± 1. (Note of course that $(0,0,0)$ does not lie on $\mathcal{S}$.)

So in summary, there are **26 LM-critical points** given by the following list:

- $\left(\pm\frac{1}{\sqrt{3}}, \pm\frac{1}{\sqrt{3}}, \pm\frac{1}{\sqrt{3}}\right)$; there are 8 points in this case. The f-values are all $\frac{1}{3}$.
- $\left(\pm\frac{1}{\sqrt{2}}, \pm\frac{1}{\sqrt{2}}, 0\right)$; there are 4 points in this case. The f-values are all $\frac{1}{2}$.
- $\left(\pm\frac{1}{\sqrt{2}}, 0, \pm\frac{1}{\sqrt{2}}\right)$; there are 4 points in this case. The f-values are all $\frac{1}{2}$.
- $\left(0, \pm\frac{1}{\sqrt{2}}, \pm\frac{1}{\sqrt{2}}\right)$; there are 4 points in this case. The f-values are all $\frac{1}{2}$.
- $(\pm 1, 0, 0)$; there are 2 points in this case. The f-values are all 1.
- $(0, \pm 1, 0)$; there are 2 points in this case. The f-values are all 1.
- $(0, 0 \pm 1)$; there are 2 points in this case. The f-values are all 1.

Phew! Okay. The other cases are much shorter:

2. $\mathcal{S}$ has no boundary to consider.
3. $\mathcal{S}$ has no limit cases to consider.

3. $\mathcal{R}$ has no limit cases to consider.

Okay, marathon done. Collate everything together. The values of f we saw were 0, $\frac{1}{3}$, $\frac{1}{2}$ and 1, and there were no limit cases of any sort. So:

- $f(0,0,0) = 0$ is the global minimum.
- $f(\pm 1, 0, 0) = f(0, \pm 1, 0) = f(0, 0, \pm 1) = 1$ are the global maximums. □

§18.5 [SIDENOTE] A little common sense can you save you a lot of work

If you step back and think a bit before you try to dive into calculus, you might find that having a bit of “common sense” might save you a lot of work. What I mean is, imagine you gave the question to your high school self before you learned *any* calculus at all. Would they be able to say anything about what properties the answer could have? The answer is, yes, pretty often.

Let’s take the example we just did: we asked for the minimum and maximum of

$$f(x, y, z) = x^4 + y^4 + z^4$$

over the region $x^2 + y^2 + z^2 \le 1$. To show the recipe, I turned off my brain and jumped straight into a really long calculation. But it turns out you can cut out a lot of the steps if you just use some common sense, not involving any calculus:

- The *minimum* is actually obvious: it’s just 0, because fourth powers are always nonnegative! So $f \ge 0$ is obvious even to a high schooler, and $f(0,0,0) = 0$.
- The *maximum* is not obvious, but actually you can see *a priori* that it must occur on the boundary $x^2 + y^2 + z^2 = 1$. Why is this? Suppose you had a point in the strict interior $P = (0.1, 0.2, 0.3)$ with $f > 0$. Then $f(P) = f(0.1, 0.2, 0.3)$ is some number. But you could obviously increase the value of f just by scaling the absolute value of things in P! For example, if I double all the coordinates of P to get $Q = (0.2, 0.4, 0.6)$, then $f(Q) = 16f(P)$. As long as Q stays within the sphere, this will be a better value.

  So any point in the interior is obviously not a maximum: if you have a point P strictly the interior, you could increase $f(P)$ by changing P to have absolute value.

That means if we had used a bit of common sense, we could have gotten the minimum with no work at all, and we could have skipped straight to the LM step for the maximum. So if you aren’t too overwhelmed by material already in this class, look for shortcuts like this when you can.

§18.6 [SIDENOTE] Compactness as a way to check your work

This is an optional section containing a nice theorem from 18.100 that could help you check your work, but isn't necessary in theory if you never make mistakes. (But in practice…)

I need a new word called "compact", and like before, it's beyond the scope of 18.02 to give a proper definition. However, I will hazard the following one: for 18.02 examples, $\mathcal{R}$ **is compact if there are no limit cases**. That is,

- All the constraints are $=$, $\leq$, or $\geq$; no $<$ or $>$,
- None of the variables can go to $\pm\infty$.

> **Tip: Compact optimization theorem**
>
> If $\mathcal{R}$ is a compact region, and f is a function to optimize on the region which is continuously defined everywhere, then there must be at least one global minimum, and at least one global maximum.

This works in both the easy case (no Lagrange multipliers) and the hard case (with Lagrange multipliers).

> **Example**
>
> Here's some examples of how this theorem can help you:
>
> - Suppose you are asked to optimize a continuous function $f(x, y)$ on the square $-1 \leq x \leq 1$, $-1 \leq y \leq 1$. We saw this square has no limit cases. Then the compact optimization theorem promises you that the answer "no global minimum" or "no global maximum" will never occur.
> - Suppose you are asked to optimize a continuous function $f(x, y, z)$ on the sphere $x^2 + y^2 + z^2 = 1$ (which means you are probably going to use Lagrange multipliers). We saw this sphere has no limit cases (and not even a boundary). Then the compact optimization theorem promises you that the answer "no global minimum" or "no global maximum" will never occur.
> - Suppose you are asked to optimize a continuous function $f(x, y, z)$ on the closed ball $x^2 + y^2 + z^2 \leq 1$, like in the last example. This closed ball also has no limit cases, so the compact optimization theorem promises you that the answer "no global minimum" or "no global maximum" will never occur.

§18.7 [RECAP] Recap of Part Foxtrot on Optimization

- We introduced the notion of critical points as points where $\nabla f = \mathbf{0}$.
  - We saw that critical points could be local minimums, local maximums, or saddle points.
  - We introduced the second derivative test as a way to tell some of these cases apart, although the second derivative test can be inconclusive.
- We talked about how regions have dimensions, limit cases, and boundaries. Although we didn't give a proper definition, we explain rules of thumb that work in 18.02.
- For optimization problems with no $=$ constraints, we check the critical points, limit cases, and boundaries.
- For optimization problems with one $=$ constraints, we check the LM-critical points, limit cases, and boundaries.

§18.8 [EXER] Exercises

TODO: Make some up

Part Golf: Multivariate integrals

For comparison, this corresponds to §13 and §17 of Poonen's notes.

§19 A zoomed out pep talk of Part Golf

This whole section is a pep talk. We'll get to recipes and details in subsequent sections.

§19.1 [TEXT] The big table of integrals

The rest of 18.02 is going to cover a bunch of different integrals. If you've been following my advice to pay attention to type safety so far, it'll help you here. I'll freely admit that I (Evan) often make type-errors in this part of 18.02 as well, so don't let your guard down.

Remember that:

> **Idea**
>
> Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is given, and $0 \le d \le n$. The point of a d-dimensional integral of f is to add up all the values of f among some d-dimensional object living in $\mathbb{R}^n$.

For example, this idea even makes sense for $d = 0$! In 18.02, a 0-dimensional object is a point (or a bunch of points), and you can evaluate f at a point by just plugging it in. So philosophically, a 0-dimensional integral is just a finite sum of f at some points. This might seem stupid that I bring up this degenerate case, but it turns out later when we cover div/grad/curl the 0-dimensional case is relevant.

With that, I present to you the following chart of ten different kinds of integrals, one for each (d, n) with $0 \le d \le n \le 3$. See Figure 20 in all its glory. (The chart is so big it doesn't quite fit in the page, but you can download a large PDF version).

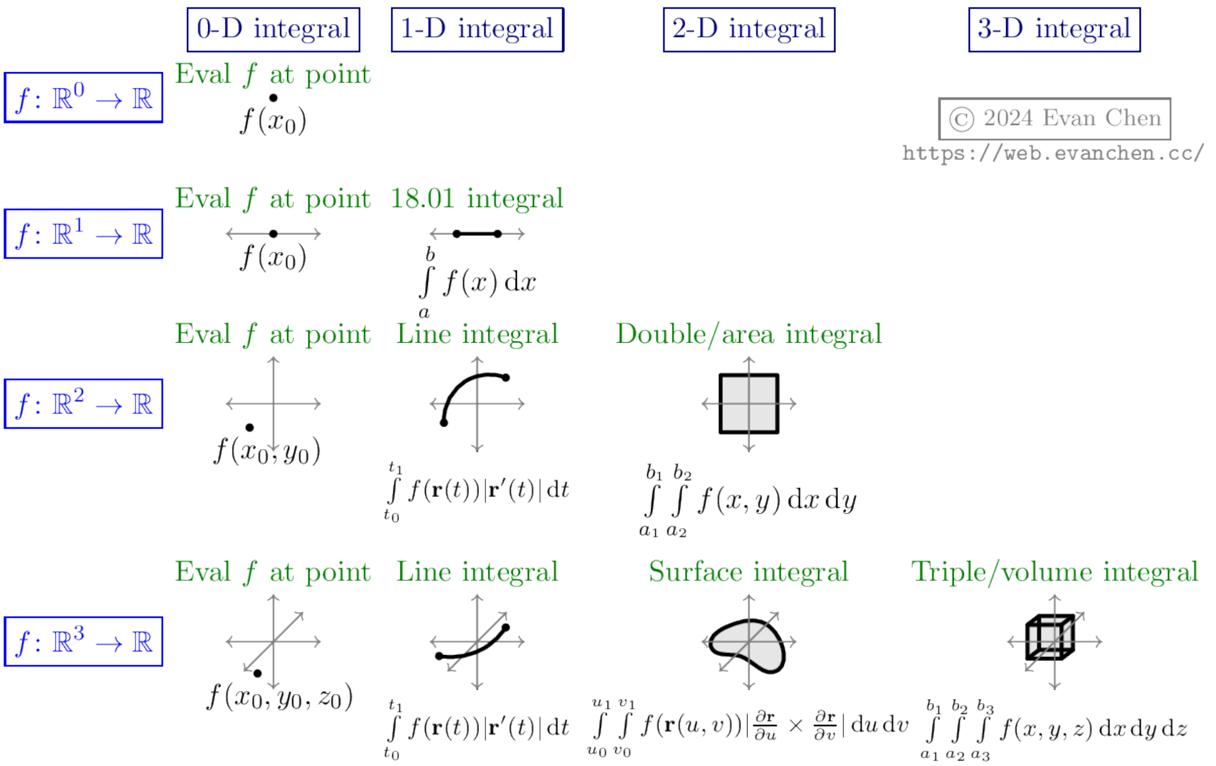

Figure 20: For each $0 \le d \le n \le 3$, the kind of integral is drawn and named. Download at https://web.evanchen.cc/upload/1802/integrals-triangles.pdf.

Here's a rundown of the things in the chart.

- The case $d = 0$ is stupid, as I just said, and it's only here because I'll reference it later.
- The case $d = 1$ and $n = 1$ was covered in 18.01. Good old single-variable integral computed using the antiderivative, via the fundamental theorem of calculus.
- After that, the conceptually simplest cases are actually $d = n = 2$ and $d = n = 3$ — the ones on the diagonal. In general, these might be called **double/area integrals** for $n = 2$ and **triple/surface integrals** for $n = 3$. We'll say a bit in a moment about how to compute these in practice, but the good news is that often you can just chain together old 18.01 integrals; you don't even need a parametrization some of the time.
- When $d = 1$ and $2 \le n \le 3$, what you get are **line integrals**. The idea is that you have a trajectory in $\mathbb{R}^n$ which is defined by some parametric equation $\mathbf{r} : \mathbb{R} \to \mathbb{R}^n$. You also have a function $f : \mathbb{R}^n \to \mathbb{R}$. The line integral lets you add up the values of f along the trajectory.

  This is just turns out to be a *single* 18.01 integral. Usually your path is parametrized by a single variable t. So even though the expression inside the integral

  $$\int_{t_0}^{t_1} f(\mathbf{r}(\mathbf{t}))\ |\mathbf{r}'(t)|\,\mathrm{d}t$$

  inside the integral might look intimidating, if you are really given a concrete f and $\mathbf{r}(t)$, then what you *really* have is

  $$\int_{t_0}^{t_1} [\text{expression involving only } t]\,\mathrm{d}t$$

  which is an 18.01 integral! And so that's something you already know how to do.

  In other words, if you have $d = 1$ and $n > 1$, you basically replace it right away with a single integral over the parametrizing line segment. In other words **line integrals translate directly into single 18.01 integrals**
- When $d = 2$ and $n = 3$, we have the **surface integral**. To compute these, you usually have to parametrize a *surface*; but since a surface is two-dimensional, rather than $\mathbf{r}(t)$ for a time parameter t you have $\mathbf{r}(u, v)$ for two parameters u and v to describe the surface. That makes these a little more annoying.

  But like the line integral, after you work out the parametrization stuff, the surface integral will transform into a 2-variable area integral. In other words **surface integrals translate directly into area integral**.

So the bottom trio — 2D/3D line integral and surface integral — end up being special instances of the single and double integrals. We'll see some examples of this later; but it'll actually be the *last* thing we cover in part Golf. Most of part Golf will be dedicated towards double and triple integrals instead.

§19.2 [TEXT] Idea of how these are computed when $d = n$ and $n \ge 2$

So as I just said, focus for now on $d = n = 2$ or $d = n = 3$ (the double and triple integral cells in chart Figure 20).

The easiest cases are when the region you're integrating is a rectangle or prism. Despite looking scary because of the number of integral signs, they are actually considered the "easy case" to think about for practical calculations:

- A double integral over a rectangle is two 18.01 integrals followed one after another.
- A triple integral over a rectangular prism really is three 18.01 integrals followed one after another.

Then there are cases where $d = n = 2$ or $d = n = 3$ but the region is not rectangular. For example, maybe in $\mathbb{R}^2$ you are trying to do an **area integral** over the disk $x^2 + y^2 \le 1$ or you are trying to do a **volume integral** over the ball $x^2 + y^2 + z^2 \le 1$ for example.

- Even in this case, sometimes you could still set up a double integral or triple integral without having to change variables. For example, an integral over the disk

$$\iint_{x^2+y^2\le 1} f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

  might actually be rewritten a double integral

$$\int_{-1}^{1}\int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} f(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

  Although it looks more frightening because the limits of integration are expressions and not numbers, it doesn't require any new techniques. It really is just two 18.01 integrals, one after another.
- If rewriting as a double or triple integral fails, then the strategy is instead to **change variables**. This method will be covered extensively later.

So to summarize

Idea

Whenever you try to compute a multivariable integral in Figure 20, your goal is to translate it into a rectangular-looking single/double/triple integral, then evaluate by using your old 18.01 methods multiple times.

This is actually really, really good news! You might have remembered from 18.01 that computing integrals of single-variable functions like $\int e^x \sin(x)$ was, well, hard![12] Computing antiderivatives was not easy at all; in fact, it's so nontrivial that MIT students made an event called the integration bee that's like the spelling bee but for integrals (I'm not kidding). You might have feared that in 18.02, you might need to learn something even more horrifying.

But no, you don't! It's a lot like how you might have been scared of multivariate differentiation at first, with the symbols ∇f or partial derivatives, until you realize that calculating partial derivatives is something *actually already know how to do* from 18.01.

The same will be true for multivariable integrals. The challenge won't actually be the anti-derivatives, which are unchanged for 18.01. The hard part will actually be figuring out the *limits* of integration!

[12]It's $\frac{1}{2}e^x(\sin(x) - \cos(x)) + C$, by the way.

§20 Double and triple integrals

§21 Change of variables

Part Hotel: Grad, Curl, and Div

For comparison, this corresponds to §14, §15, §18, §19, §20, §21 of Poonen's notes.

§22 Vector fields

§22.1 [TEXT] Vector fields

In Part Golf, we only considered integrals of scalar-valued functions. However, in Part Hotel we will meet a **vector field**, which is another name for a function that inputs points and outputs vectors.

> **Definition**
>
> A **vector field** for $\mathbb{R}^n$ is a function $F : \mathbb{R}^n \to \mathbb{R}^n$ that assigns to each point $P \in \mathbb{R}^n$ a vector $\mathbf{F}(P) \in \mathbb{R}^n$.

You actually have met these before

> **Example**
>
> Every gradient is an examples of a vector fields! That is, if $f : \mathbb{R}^n \to \mathbb{R}$, then ∇f is a vector field for $\mathbb{R}^n$.

In fact, there's a word for this:

> **Definition**
>
> A vector field for $\mathbb{R}^n$ is called **conservative** if it happens to equal ∇f for some function $f : \mathbb{R}^n \to \mathbb{R}$.

In Part Hotel we'll meet vector fields that aren't conservative too.

> **Type signature**
>
> For standalone vector fields, we'll always use capital bold letters like $\mathbf{F}$ to denote them. That said, remember ∇f is *also* a vector field. So that's why the operator ∇ itself is typeset a little bit bold.
>
> Like the gradient, you should draw inputs to $\mathbf{F}$ as *points* (dots) but the outputs as *vectors* (arrows). Don't mix them.

§22.2 [TEXT] How do we picture a vector field?

There's a lot of ways to picture a vector field, especially in physics. For consistency, I'm going to pick *one* such framework and write all my examples in terms of it. So **in my book, all examples will be aquatic** in nature; but if you can't swim[13], you should feel free to substitute your own. Imagine an electric field. Or a black hole in outer space. Or air currents in the atmosphere. Whatever works for you!

Anyway, for my book, we'll use the following picture:

[13] Doesn't MIT make you pass a swim test, though?

Idea

Imagine a flowing body of water (ocean, river, whirlpool, fountain, etc.) in $\mathbb{R}^2$ or $\mathbb{R}^3$. Then at any point, we draw a tiny arrow $\mathbf{F}(P)$ indicating the direction and speed of the water at the point P. You could imagine if you put a little ball at the point P, the current would move the ball along that arrow.

Sounds a lot like the gradient, right? Indeed, conservative vector fields are a big family of vector fields, and so we should expect they fit this picture pretty neatly. But the thing about conservative vector fields is this: ∇f, as a vector field, is always rushing *towards* whatever makes the value of F bigger. Whereas generic vector fields might, for example, go in loops. Let's put these examples into aquatic terms.

Example of a conservative vector field: going downstream a river

Let's imagine we have a river with a strong current. We'll make the important assumption that the river only goes one way: that is, if you go along the current, you never end up back where you started. In real life, this often occurs if the river goes down a mountain, so as you go down the river you're losing elevation.

If you do this, you can define a "downstream function" $f : \mathbb{R}^3 \to \mathbb{R}$ as follows: for every point P in the river, $f(P)$ measures how far downstream you are. For example, if the river had a head, maybe we could assign f the value zero there, and then f would increase as you get farther from the bank, reaching the largest value at the mouth. (For mountainous rivers, f might instead be thought of as decreasing in elevation.)

Then the vector field corresponding to the river is the gradient ∇f. Remember, the gradient of f tells you what direction to move in to increase f. And if you throw a ball into a river, its motion could be described simply as: the ball moves downstream.

TODO: draw a picture of a river

Example of a non-conservative vector field: a whirlpool

Now imagine instead you have a whirlpool. If you throw a ball in it, it goes in circles around vertex of the whirlpool. This doesn't look anything like the river! If you have a river, you never expect a ball to come back to the same point after a while, because it's trying to go downstream. But with a whirlpool, you keep going in circles over and over.

If you draw the vector field corresponding to a whirlpool, it looks like lots of concentric rings made by tiny arrows. That's an example of a non-conservative vector field.

TODO: draw a picture of a whirlpool

§22.3 [TEXT] Preview of integration over vector fields

So far everything's great. But soon we'll have to start integrating over vector fields. That's when the type signatures go crazy.

In order for this to be even remotely memorable, what I'm going to do is augment the previous Figure 20 with pictures corresponding to the situations in which we might integrate a vector field. The new chart can also be downloaded as a large PDF version.

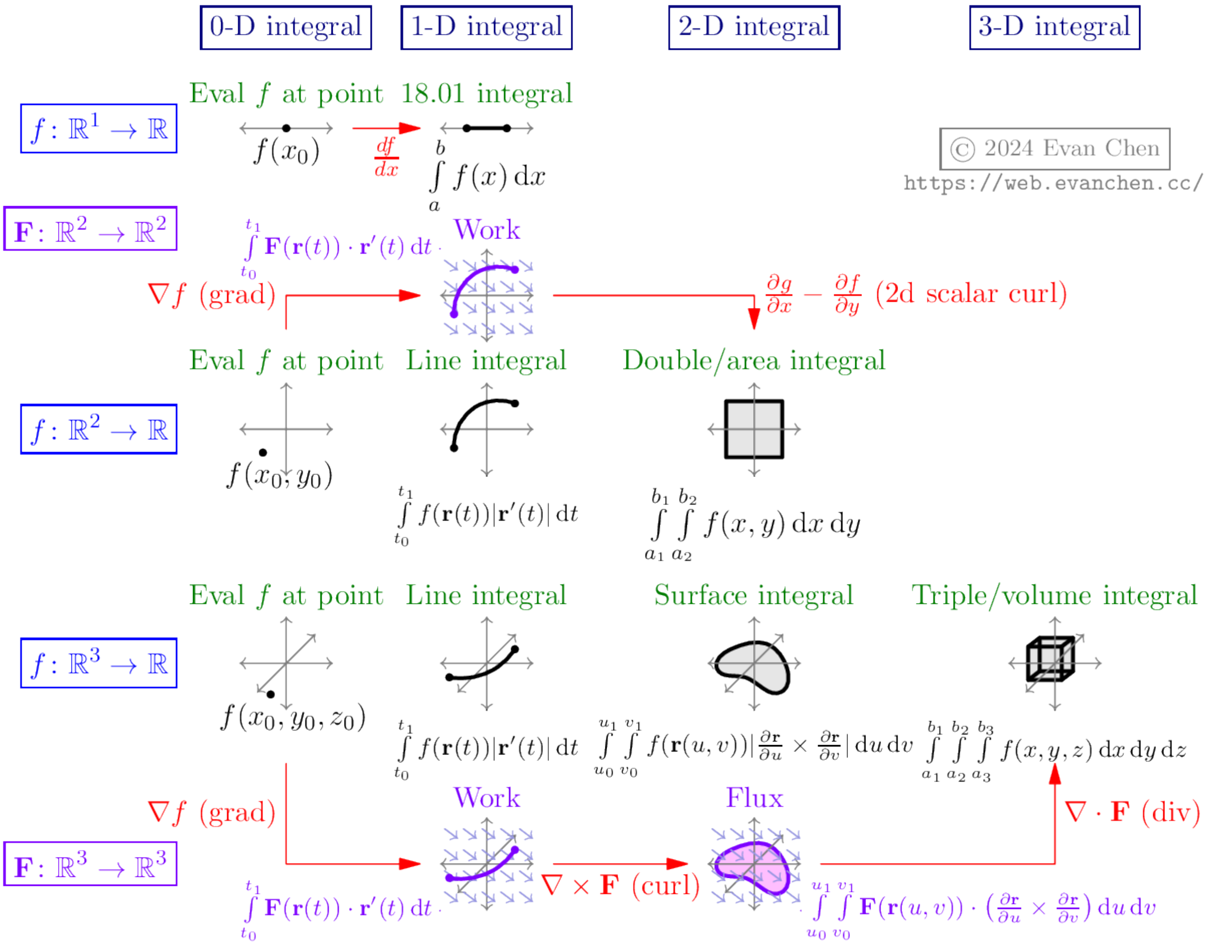

Figure 21: Upgraded Figure 20 with vector fields. Download at `https://web.evanchen.cc/upload/1802/integrals-stokes.pdf`.

There are two new features of Figure 21 compared to the old version: the three purple pictures and the six red arrows. We'll define them all over the next few sections, so just a few words now.

§22.3.1 The three purple pictures

There are **three new pictures in purple**: they are **work** (for 1-D case) and **flux** (2-D case). Basically, these are the only two situations in which we'll be integrating over a vector field:

- either we have a path along a vector field and want to measure the *work* of the vector field *along* that path (in the physics sense),
- or we have a surface in a 3-D vector field and want to measure the *flux* of the vector field *through* the surface.

These terms will be defined next section.

</> Type signature

The new purple things are *still* all scalar quantities, i.e. work and flux are both numbers, not vectors.

§22.3.2 The six red arrows

There are also **six new red arrows**. They indicate transformations on functions: a way to take one type of function and use it to build another function.

For example, the gradient ∇ is the one we've discussed: if you start with a scalar-valued function $f : \mathbb{R}^n \to \mathbb{R}$, the gradient creates into a vector field $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$. (The $\frac{\mathrm{d}f}{\mathrm{d}x}$ in the $f : \mathbb{R}^1 \to \mathbb{R}$ case is also just the gradient, though a bit more degenerate.)

We'll soon meet three more transformations:

- **2D curl**, which converts a vector field on $\mathbb{R}^2$ *back* into a scalar-valued function;
- **3D curl**, which converts a vector field on $\mathbb{R}^3$ into *another* vector field;
- **divergence**, which converts a vector field on $\mathbb{R}^3$ *back* into a scalar-valued function.

§22.3.3 Stay determined

This is probably super overwhelming right now, and Figure 21 might be frightening to look at because there's so much information in it. Don't worry, we'll take Figure 21 apart one piece at a time over the rest of the semester. This will be a three-phase program:

- First, I'll tell you how to *integrate* work and flux of a vector field.
- Second, I will define *individually* each of the three transformations grad, curl, div. (Actually we defined the first one already, so it's just curl and div.)
- Third, I'll tell you how grad, curl, and div interact with each other, using the notorious *generalized Stokes' theorem*.

§22.4 [EXER] Exercises

Exercise 22.1: Take a few deep breaths, touch some grass, and have a nice drink of water, so that you can look at Figure 21 without feeling fear.

Exercise 22.2: Print out a copy of the high-resolution version of Figure 21 (which can be downloaded at `https://web.evanchen.cc/upload/1802/integrals-stokes.pdf`) and hang it in your room.

§23 Work and flux

§24 Grad, curl, and div, individually

§25 Generalized Stokes' theorem

§25.1 [TEXT] The only two things you need to remember for this section

Remember the red arrows in Figure 21? If you followed my advice in Exercise 22.2, you probably remember where the red arrows in the picture are now. Now it'll pay off in spaced, because there's only two things you need to know about them for this section.

Memorize: Two red arrows gives zero

In Figure 21, if you follow two red arrows consecutively, you get zero.

Memorize: Generalized Stokes' Theorem, for 18.02

In Figure 21, take any of the six red arrows

$$X \to \operatorname{del}(X).$$

Let $\mathcal{R}$ be a compact region. Then the integral of X over the **boundary** of $\mathcal{R}$ equals the integral of $\operatorname{del}(X)$ over $\mathcal{R}$:

$$\int_{\text{boundary}(\mathcal{R})} X = \int_{\mathcal{R}} \operatorname{del}(X).$$

Because the chart in Figure 21 is so big, the first item will give 3 different theorems, while the second item will give 6 different theorems (one for each red arrow). But you don't need to memorize all $3 + 6 = 9$ results. All you have to do is remember the two items above. Then all 9 results will fall out.

Solutions to the exercises and problems

§26 Solutions to the challenge problems from midterm 1

The problem statements are given in Section 11.

§26.1 Solution to Problem 11.1

Answer: $\begin{pmatrix} \frac{1}{2} \\ \frac{3}{2} \\ -1 \end{pmatrix}$.

§26.1.1 First approach using vector projection

In recitation R02 you had to calculate the distance from a vector to a plane. This problem only requires one step on top of that: you need to then translate by the normal vector. See the cartoon below, where $\mathbf{a}$ denotes the answer.

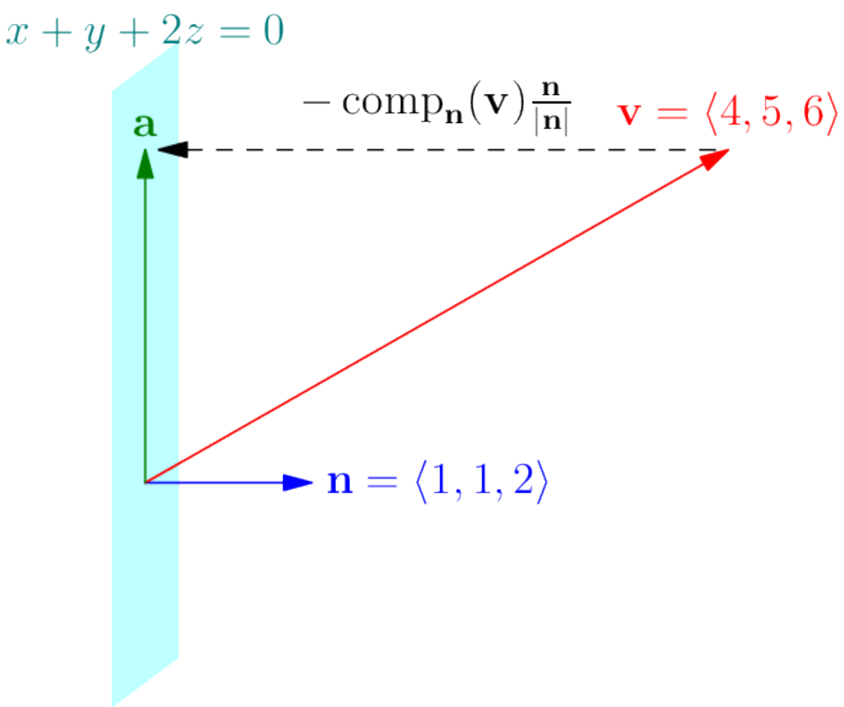

Figure 22: Projection onto a plane.

To execute the calculation, let $\mathbf{v} = \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix}$ and $\mathbf{n} = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}$. The scalar projection is

$$\operatorname{comp}_{\mathbf{n}}(\mathbf{v}) = \frac{\mathbf{v} \cdot \mathbf{n}}{|\mathbf{n}|} = \frac{21}{\sqrt{6}}.$$

The vector projection is then

$$(\operatorname{comp}_{\mathbf{n}}(\mathbf{v}))\frac{\mathbf{n}}{|\mathbf{n}|} = \frac{21}{\sqrt{6}} \frac{\begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}}{\sqrt{6}} = \begin{pmatrix} \frac{7}{2} \\ \frac{7}{2} \\ 7 \end{pmatrix}.$$

Then the desired projection is

$$\mathbf{v} - \operatorname{proj}_{\mathbf{n}}(\mathbf{v}) = \begin{pmatrix} \frac{1}{2} \\ \frac{3}{2} \\ -1 \end{pmatrix}.$$

§26.1.2 Second approach using normal vectors only (no projection stuff)

A lot of you don't find vector projection natural (I certainly don't). So it might be easier to imagine shifting $\mathbf{v}$ by *some* multiple of $\mathbf{n} = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}$ and then work out which multiple it is.

Specifically, we're looking for[14] a real number $t \in \mathbb{R}$ such that the vector

$$\mathbf{a} = \mathbf{v} - t\mathbf{n} = \begin{pmatrix} 4-t \\ 5-t \\ 6-2t \end{pmatrix}$$

lies on the plane $x + y + 2z = 0$. But we can actually solve for t just by plugging this $\mathbf{a}$ into the equation of the plane:

$$(4-t) + (5-t) + 2(6-2t) = 0 \Longrightarrow 21 - 6t = 0 \Longrightarrow t = \frac{7}{2}.$$

Hence the answer

$$\mathbf{a} = \begin{pmatrix} 4 - \frac{7}{2} \\ 5 - \frac{7}{2} \\ 6 - 2\left(\frac{7}{2}\right) \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ \frac{3}{2} \\ -1 \end{pmatrix}.$$

[14]In comparison to the first solution, the value of t is exactly

$$t = \frac{\operatorname{comp}_{\mathbf{n}}(\mathbf{v})}{|\mathbf{n}|}.$$

But the idea behind the second solution is that you don't *need to know* what the geometric formula of t is. You can just solve for t indirectly by asserting that $\mathbf{a}$ lies on $x + y + 2z = 0$.

§26.2 Solution to Problem 11.2

Answer: This equals the volume of the parallelepiped formed by $\overrightarrow{DA}$, $\overrightarrow{DB}$, $\overrightarrow{DC}$.

Here are two approaches for proving it.

§26.2.1 First approach using coordinates

Let $D = (0, 0, 0)$, $A = (x_A, y_A, b_A)$, $B = (x_B, y_B, z_B)$, $C = (x_C, y_C, z_C)$. Then expanding the cross product gives

$$(x_A\mathbf{e}_1 + y_A\mathbf{e}_2 + z_A\mathbf{e_3}) \cdot \det\left(\begin{pmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \\ x_B & y_B & z_B \\ x_C & y_C & z_C \end{pmatrix}\right).$$

If you think about what evaluating the determinant using the formula together with the dot product would give, you should find it's actually just

$$\det\left(\begin{pmatrix} x_A & y_A & z_A \\ x_B & y_B & z_B \\ x_C & y_C & z_C \end{pmatrix}\right)$$

which is the volume of the parallelepiped.

§26.2.2 Second approach using geometric picture

The cross product $\overrightarrow{DB} \times \overrightarrow{DC}$ is a vector whose area is equal to the parallelogram formed by $\overrightarrow{DB}$ and $\overrightarrow{DC}$. The dot product of that cross product against $\overrightarrow{DA}$ is equal to the *height* of A to plane BCD times this area, and the volume is the height times the area. See the following picture from https://en.wikipedia.org/wiki/Triple_product (in the Wikipedia figure, $\mathbf{a}$ denotes our $\overrightarrow{DA}$, etc.).

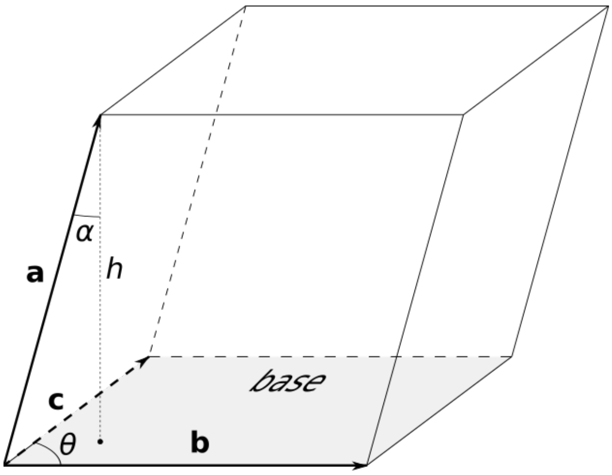

Figure 23: Triple product image taken from Wikipedia.

§26.3 Solution to Problem 11.3

Answer: 0, no matter which plane $\mathcal{P}$ is picked.

§26.3.1 First approach using basis vectors

Let $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$ be the three basis vectors. Then:

- The matrix M is formed by gluing $f(\mathbf{e}_1)$, $f(\mathbf{e}_2)$, $f(\mathbf{e}_3)$ together.
- I claim the vectors $f(\mathbf{e}_1)$, $f(\mathbf{e}_2)$, $f(\mathbf{e}_3)$ are linearly dependent. After all, they are all contained in the two-dimensional plane $\mathcal{P}$ by definition, and so three vectors in a plane can't be linearly independent.
- So the determinant is equal to zero (this theorem is one of the criteria we use to check whether vectors are linearly independent or not).

§26.3.2 Second approach using eigenvectors

Let $\mathbf{n}$ be any nonzero normal vector to $\mathcal{P}$. Then $f(\mathbf{n}) = \mathbf{0}$, so $\mathbf{n}$ is an eigenvector with eigenvalue 0. Since the determinant is the product of the eigenvalues, the determinant must be 0 too.

§26.3.3 Third approach using coordinate change

This approach requires you to know the fact that the determinant doesn't change if you rewrite the matrices in a new basis.

Let $\mathbf{n}$ be any nonzero normal vector to $\mathcal{P}$. Pick two more unit vectors $\mathbf{b}_1$ and $\mathbf{b}_2$ perpendicular to $\mathbf{n}$ that span $\mathcal{P}$. Then $\mathbf{b}_1$, $\mathbf{b}_2$ and $\mathbf{n}$ are linearly independent and spanning, i.e. a basis of $\mathbb{R}^3$. So we can change coordinates to use these instead.

We know that

$$\begin{aligned} M(\mathbf{b}_1) &= \mathbf{b}_1 \\ M(\mathbf{b}_2) &= \mathbf{b}_2 \\ M(\mathbf{n}_2) &= \mathbf{0}. \end{aligned}$$

If we wrote M as a matrix *in this new basis* $\langle \mathbf{b}_1, \mathbf{b}_2, \mathbf{n} \rangle$ (rather than the usual basis), we would get the matrix

$$M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

which has determinant 0.

> *i* Remark
>
> In fact, if you also know that the trace doesn't change when you rewrite M in a different basis, this approach shows the trace M is always exactly $1 + 1 + 0 = 2$ as well, no matter which plane $\mathcal{P}$ is picked.

§26.4 Solution to Problem 11.4

Answer: $|\mathbf{a} \times \mathbf{v}| = 3$ and $|\mathbf{b} \times \mathbf{v}| = 2$.

Since $\mathbf{v}$ is contained in the span of $\mathbf{a}$ and $\mathbf{b}$, we can just pay attention to the plane spanned by these two perpendicular unit vectors. So the geometric picture is that $\mathbf{v}$ can be drawn in a rectangle with $\mathbf{a}$ and $\mathbf{b}$ as a basis, as shown. Because $\mathbf{v} \cdot \mathbf{a} = 2$ and $\mathbf{v} \cdot \mathbf{b} = 3$, this rectangle is 2 by 3.

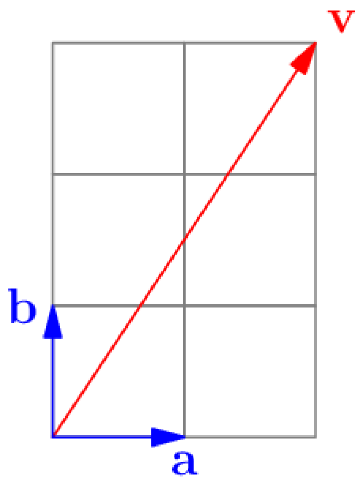

Figure 24: Plotting $\mathbf{v}$ in the span of $\mathbf{a}$ and $\mathbf{b}$.

Now the magnitude of the cross product $\mathbf{a} \times \mathbf{v}$ is supposed to be equal to the area of the parallelogram formed by $\mathbf{a}$ and $\mathbf{v}$, which is 3 (because this parallelogram has base $|\mathbf{a}| = 1$ and height $|\mathbf{v} \cdot \mathbf{b}| = 3$). Similarly, $\mathbf{b} \times \mathbf{v}$ has magnitude 2.

§26.5 Solution to Problem 11.5

Answer: 0.

There are several approaches possible. The first two show how to find the four entries of the matrix M. The latter sidestep this entirely and show that the matrix is actually always trace 0.

§26.5.1 First approach: brute force

Like in the pop quiz in my R04 notes, we will try to work out $M\begin{pmatrix}1\\0\end{pmatrix}$ and $M\begin{pmatrix}0\\1\end{pmatrix}$. We're looking for constants c_1 and c_2 such that $c_1\begin{pmatrix}4\\7\end{pmatrix} + c_2\begin{pmatrix}5\\9\end{pmatrix} = \begin{pmatrix}1\\0\end{pmatrix}$.

- Solving the system of equations $4c_1 + 5c_2 = 1$ and $7c_1 + 9c_2 = 0$ using your favorite method gives coefficients $c_1 = 9$ and $c_2 = -7$, i.e.

$$9\begin{pmatrix}4\\7\end{pmatrix} - 7\begin{pmatrix}5\\9\end{pmatrix} = \begin{pmatrix}1\\0\end{pmatrix}.$$

  This lets us get

$$M\left(\begin{pmatrix}1\\0\end{pmatrix}\right) = 9M\left(\begin{pmatrix}4\\7\end{pmatrix}\right) - 7M\left(\begin{pmatrix}5\\9\end{pmatrix}\right) = 9\begin{pmatrix}5\\9\end{pmatrix} - 7\begin{pmatrix}4\\7\end{pmatrix} = \begin{pmatrix}17\\32\end{pmatrix}.$$

- By solving the analogous system we can find the identity

$$-5\begin{pmatrix}4\\7\end{pmatrix} + 4\begin{pmatrix}5\\9\end{pmatrix} = \begin{pmatrix}0\\1\end{pmatrix},$$

  and hence:

$$M\left(\begin{pmatrix}0\\1\end{pmatrix}\right) = -5M\left(\begin{pmatrix}4\\7\end{pmatrix}\right) + 4M\left(\begin{pmatrix}5\\9\end{pmatrix}\right) = -5\begin{pmatrix}5\\9\end{pmatrix} + 4\begin{pmatrix}4\\7\end{pmatrix} = \begin{pmatrix}-9\\-17\end{pmatrix}.$$

Gluing these together

$$M = \begin{pmatrix}17 & -9\\32 & -17\end{pmatrix}.$$

The trace is thus $17 + (-17) = 0$.

§26.5.2 Second approach: inverse matrices

We can collate the two given equations into saying that

$$M\begin{pmatrix}4 & 5\\7 & 9\end{pmatrix} = \begin{pmatrix}5 & 4\\9 & 7\end{pmatrix}.$$

Hence one could also recover M by multiplying by the inverse matrix:

$$M = \begin{pmatrix}5 & 4\\9 & 7\end{pmatrix}\begin{pmatrix}4 & 5\\7 & 9\end{pmatrix}^{-1} = \begin{pmatrix}5 & 4\\9 & 7\end{pmatrix}\frac{1}{4\cdot 9 - 7\cdot 5}\begin{pmatrix}9 & -5\\-7 & 4\end{pmatrix} = \begin{pmatrix}17 & -9\\32 & -17\end{pmatrix}.$$

(Of course, we get the same entries for M as the last approach.) Again the trace is $17 + (-17) = 0$.

§26.5.3 Third approach: Guessing eigenvectors and eigenvalues

Let $\mathbf{b}_1 = \begin{pmatrix}4\\7\end{pmatrix}$ and $\mathbf{b}_2 = \begin{pmatrix}5\\9\end{pmatrix}$. Adding and subtracting the given equations gives

$$\begin{aligned} M(\mathbf{b}_1 + \mathbf{b}_2) &= \mathbf{b}_1 + \mathbf{b}_2 \\ M(\mathbf{b}_1 - \mathbf{b}_2) &= -(\mathbf{b}_1 - \mathbf{b}_2). \end{aligned}$$

So $\mathbf{b}_1 \pm \mathbf{b}_2$ are eigenvectors with eigenvalues ± 1. Since M is a 2×2 matrix there are at most two eigenvalues: we found them all!

The trace of M is the sum of the eigenvalues. Call in the answer $1 + (-1) = 0$.

§26.5.4 Fourth approach: Change coordinates

This approach requires you to know the fact that the trace doesn't change if you rewrite the matrices in a new basis.

Since $\mathbf{b}_1 = \begin{pmatrix} 4 \\ 7 \end{pmatrix}$ and $\mathbf{b}_2 = \begin{pmatrix} 5 \\ 9 \end{pmatrix}$ are a basis of $\mathbb{R}^2$, we can change coordinates to use the $\mathbf{b}_i$. In that case,

$$M(\mathbf{b}_1) = \mathbf{b}_2 \quad \text{and} \quad M(\mathbf{b}_2) = \mathbf{b}_1.$$

If we wrote M as a matrix *in this new basis* $\langle \mathbf{b}_1, \mathbf{b}_2 \rangle$ (rather than the usual basis), we would get the matrix

$$M = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

which has trace $0 + 0 = 0$.

§26.6 Solution to Problem 11.6

Answer: $\frac{3\sqrt{3}}{4}\sqrt[3]{61}$.

We start by converting the complex number $5+6i$ into polar form. The modulus r of $5+6i$ is:

$$r = |5+6i| = \sqrt{5^2+6^2} = \sqrt{25+36} = \sqrt{61}.$$

The argument θ is some random angle we won't use the exact value of: $\theta = \arg(5+6i) = \tan^{-1}\left(\frac{6}{5}\right)$.

Now to find the cube roots of $z^3 = 5+6i$, we use the polar form:

$$z = \sqrt[6]{61}\left(\cos\left(\frac{\theta+2k\pi}{3}\right) + i\sin\left(\frac{\theta+2k\pi}{3}\right)\right)$$

for $k = 0, 1, 2$. This gives us three roots corresponding to the different values of k.

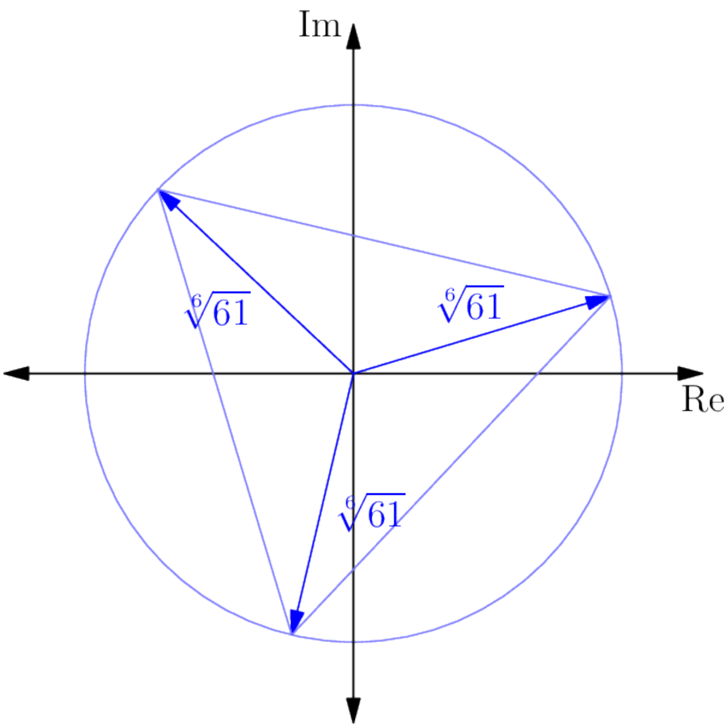

Figure 25: Three solutions to $z^3 = 5+6i$

This looks like an equilateral triangle centered around the origin, where each spoke coming from the origin has magnitude s, where

$$s = \sqrt[6]{61}.$$

If we cut up the equilateral triangle by the three arrows above, we get three small isosceles triangles with a $120°$ angle at the apex. The area of each triangle is going to be $\frac{s^2}{2}\sin(120°)$.

So this gives a final answer of

$$3 \cdot \frac{\sqrt[3]{61}}{2} \cdot \sin(120°) = \frac{3\sqrt{3}}{4}\sqrt[3]{61}.$$

Appendix

§27 Appendix

This entire section is not for exam, obviously.

§27.1 If you are thinking of majoring in math…

During the course, one of the students asked me about academic advice saying they wanted to become a math major at MIT. If that also describes you, here's what I told them. The course numbers here are with respect to MIT, but this advice should hold equally well at other universities.

§27.1.1 The two starter topics are algebra and analysis, not calculus

It may come as a surprise to you that 18.02 isn't a prerequisite, even indirectly, for most upper-division math classes (18.xyz for $x \geq 1$). The two most important areas to take in pure math are **18.100** (real analysis) and **18.701–18.702** (algebra); these are sort of the barrier between the world of pre-university math and serious math. Once you clear these two classes, the floodgates open and the world of modern math is yours to explore (see the dependency chart in the Napkin for more on this).

For example, if you take 18.701, the instructor will literally *throw away* the "definitions" of linear transformations (and others) you learned in 18.02 and replace them with the "correct" ones. You've seen me do this already. Similarly, you will have new rigorous definitions of derivatives and integrals. In some sense, 18.100 is really *redoing* all of 18.01 and 18.02 with actual proofs.

§27.1.2 Proof-writing

A prerequisite to both 18.100 (real analysis) and 18.701−18.702 (algebra) isn't any particular theory, but **proof experience**, and that's the biggest priority if you don't have that yet. (And I don't mean two-column proofs in 9th grade geometry. Two-column proofs were something made up for K-12 education and never used again.)

At MIT, I've been told in recent years there's a class called 18.090 for this. This class is new enough I don't even have any secondhand accounts, but if Poonen is on the list of instructors who developed the course, I trust him. If you're at a different school, my suggestion would be to ask any of the math professors a question along the lines of "I'd like to major in math, but I don't have proof experience yet. Which class in your department corresponds to learning proof arguments?". They should know exactly what you're talking about.

Alternatively, if you are willing to study proof-writing independently, the FAQ `https://web.evanchen.cc/faq-contest.html#C-5` on my website has some suggestions. In particular, if you're a textbook kind of person, the book I used growing up was Rotman's *Journey into Math: An Introduction to Proofs*, available at `https://store.doverpublications.com/products/9780486453064` it worked well for me. I'm sure there are other suitable books as well.

§27.1.3 The three phases of math education (from Tao's blog)

Let me put proof-writing into the bigger framework. Terence Tao, on his blog, describes a division of mathematical education into three stages. The descriptions that follows are copied verbatim from that link:

1. The "pre-rigorous" stage, in which mathematics is taught in an informal, intuitive manner, based on examples, fuzzy notions, and hand-waving. (For instance, calculus is usually first introduced in terms of slopes, areas, rates of change, and so forth.) The emphasis is more on computation than on theory.

2. The "rigorous" stage, in which one is now taught that in order to do maths "properly", one needs to work and think in a much more precise and formal manner (e.g. re-doing calculus by using epsilons and deltas all over the place). The emphasis is now primarily on theory; and one is expected to be able to comfortably manipulate abstract mathematical objects without focusing too much on what such objects actually "mean".
3. The "post-rigorous" stage, in which one has grown comfortable with all the rigorous foundations of one's chosen field, and is now ready to revisit and refine one's pre-rigorous intuition on the subject, but this time with the intuition solidly buttressed by rigorous theory. (For instance, in this stage one would be able to quickly and accurately perform computations in vector calculus by using analogies with scalar calculus, or informal and semi-rigorous use of infinitesimals, big-O notation, and so forth, and be able to convert all such calculations into a rigorous argument whenever required.) The emphasis is now on applications, intuition, and the "big picture".

These notes are still in the first stage. The introduction-to-proofs class at your school will essentially be the beginning of the second stage.

§27.2 Proof that the algebraic definition of dot product matches the geometric one

We have two definitions in play and we want to show they coincide, which makes notation awkward. So in what follows, our notation $\mathbf{u} \cdot \mathbf{v}$ will always refer to the *geometric* definition; that is $\mathbf{u} \cdot \mathbf{v} := |\mathbf{u}|\ |\mathbf{v}| \cos\theta$. And our goal is to show that it matches the algebraic definition.

We will assume that $|\mathbf{u}| = 1$ (i.e. $\mathbf{u}$ is a unit vector) so that $\mathbf{u} \cdot \mathbf{v}$ is the length of the projection of $\mathbf{v}$ onto $\mathbf{u}$. This is OK to assume because in the general case one just scales everything by $|\mathbf{u}|$.

§27.2.1 Easy special case

As a warmup, try to show that if $\mathbf{u} = \binom{a}{b}$ is any vector, then $\mathbf{u} \cdot \mathbf{e}_1 = a$. (This is easy. The projection of $\mathbf{u}$ onto $\mathbf{e}_1$ is literally a.)

§27.2.2 Main proof

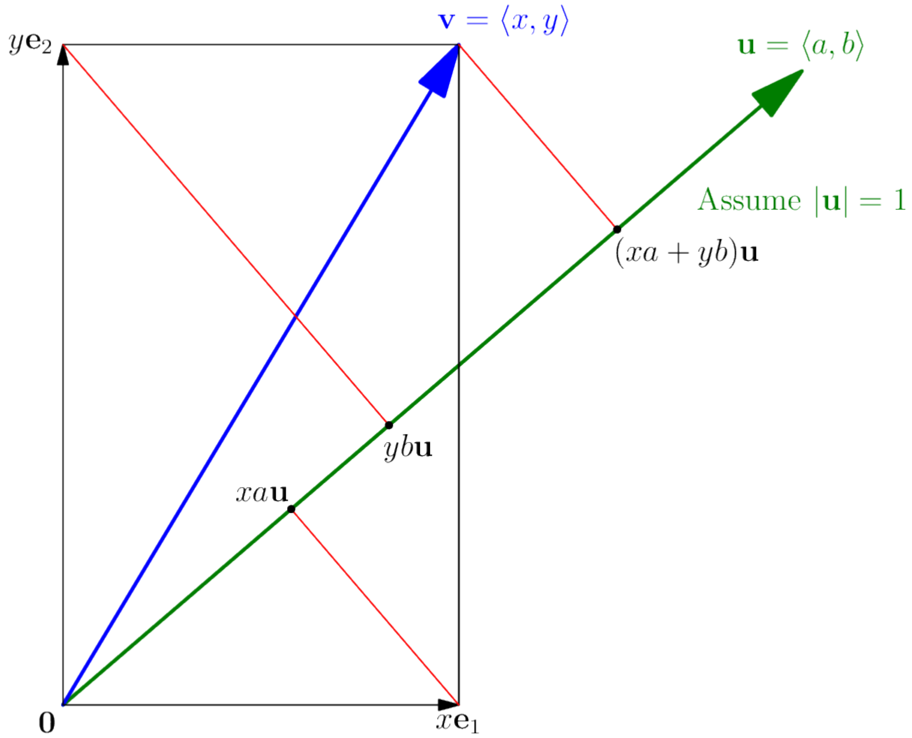

Figure 26: Proof that the dot product is given by the projection

For concreteness, specialize to $\mathbb{R}^2$ and consider $\mathbf{u} \cdot \mathbf{v}$ where $\mathbf{u} = \binom{a}{b}$ is a unit vector (i.e. $\mathbf{u} = 1$), and $\mathbf{v} = \binom{x}{y}$ is any vector in $\mathbb{R}^2$. Then we want to show that the projection of $\mathbf{v}$ onto $\mathbf{u}$ has length $xa + yb$.

The basic idea is to decompose $\mathbf{v} = x\mathbf{e}_1 + y\mathbf{e}_2$. The length of projection of $\mathbf{v}$ onto $\mathbf{u}$ can be decomposed then into the lengths of projections of $x\mathbf{e}_1$ and $y\mathbf{e}_2$. (To see this, tilt your head so the green line is horizontal; recall that the black quadrilateral is a rectangle, hence also a parallelogram). In other words,

$$\mathbf{u} \cdot \mathbf{v} = \mathbf{u} \cdot (x\mathbf{e}_1 + y\mathbf{e}_2) = x(\mathbf{u} \cdot \mathbf{e}_1) + y(\mathbf{u} \cdot \mathbf{e}_2).$$

But we already did the special cases before:

$$\begin{aligned} \mathbf{u} \cdot \mathbf{e}_1 &= a \\ \mathbf{u} \cdot \mathbf{e}_2 &= b. \end{aligned}$$

Hence, we get the right-hand side is

$$\mathbf{u} \cdot \mathbf{v} = xa + yb,$$

as advertised. In summary, by using the black parallelogram, we were able to split $\mathbf{u} \cdot \mathbf{v}$ into two easy cases we already know how to do.

The same idea will work in $\mathbb{R}^3$ if you use $\mathbf{v} = x\mathbf{e}_1 + y\mathbf{e}_2 + z\mathbf{e}_3$ instead, and replace the parallelogram with a parallelepiped, in which case one now has 3 easy cases. And so on in n dimensions.

§27.3 Saddle point simulation code for Section 16.3

```
import random

random.seed("18.02 Fall 2024")


def classify_critical_points(a3, a2, a1, b3, b2, b1):
    # f = a3 * x**3 + a2 * x**2 + a1 * x + b3 * y**3 + b2 * y**2 + b1 * y
    # the constant term has no effect on the critical points, so we ignore it
    assert a3 != 0
    assert b3 != 0

    # fx = 3 a3 x^2  + 2 a2 x + 1
    # fy = 3 b3 y^2  + 2 b2 y + 1
    # If either of these have negative discriminant, rage-quit
    if 4 * a2 * a2 - 12 * a3 * a1 < 0:
        return (0, 0, 0)
    if 4 * b2 * b2 - 12 * b3 * b1 < 0:
        return (0, 0, 0)

    # Otherwise, let's get the two critical values
    x1 = (-2 * a2 + (4 * a2 * a2 - 12 * a3 * a1) ** 0.5) / (6 * a3)
    x2 = (-2 * a2 - (4 * a2 * a2 - 12 * a3 * a1) ** 0.5) / (6 * a3)
    y1 = (-2 * b2 + (4 * b2 * b2 - 12 * b3 * b1) ** 0.5) / (6 * b3)
    y2 = (-2 * b2 - (4 * b2 * b2 - 12 * b3 * b1) ** 0.5) / (6 * b3)

    local_minima = 0
    local_maxima = 0
    saddle_points = 0

    for x0 in (x1, x2):
        for y0 in (y1, y2):
            fxx = 6 * a3 * x0 + 2 * a2
            fyy = 6 * b3 * x0 + 2 * b2
            assert fxx != 0 and fyy != 0  # give up lol
            if fxx > 0 and fyy > 0:
                local_minima += 1
            elif fxx < 0 and fyy < 0:
                local_maxima += 1
            else:
                saddle_points += 1
    return (local_minima, local_maxima, saddle_points)


local_minima = 0
local_maxima = 0
saddle_points = 0

N = 10**6
for _ in range(10000):
    a1 = random.randint(-N, N + 1)
    a2 = random.randint(-N, N + 1)
    a3 = random.randint(-N, N + 1)
    b1 = random.randint(-N, N + 1)
    b2 = random.randint(-N, N + 1)
    b3 = random.randint(-N, N + 1)
```

```
    u, v, w = classify_critical_points(a3, a2, a1, b3, b2, b1)
    local_minima += u
    local_maxima += v
    saddle_points += w
total = local_minima + local_maxima + saddle_points
print(local_minima / total, local_maxima / total, saddle_points / total, total)
```